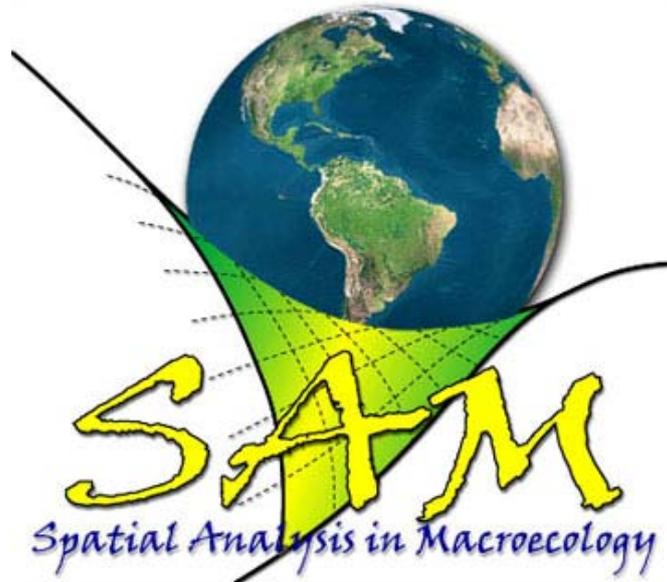




SAM

Spatial Analysis in Macroecology



SAM Tutorial

SAM workshop, IBS meeting, Crete 07 Jan 2011

Thiago Fernando L. V. B. Rangel

Department of Ecology, Federal University of Goiás, Brazil

Richard Field

School of Geography, University of Nottingham, UK

J. Alexandre F. Diniz-Filho

Department of Ecology, Federal University of Goiás, Brazil

Introduction

The SAM (“*Spatial Analysis in Macroecology*”) software was developed in the lab at Federal University of Goiás, Brasil, and aims to provide an integrated computational platform to perform spatial analyses, focused on macroecological data. The software can be obtained freely from our website www.ecoevol.ufg.br/sam, and a description of the main features of the software was published in *Global Ecology & Biogeography* (Rangel et al. 2006, GEB 15: 321-327) and in *Ecography* (Rangel et al. 2010, Ecography 33: 46-50). This short guide aims to help you to explore the program.

SAM has four main groups of modules (**File**, **Data**, **Structure** and **Modeling**), each one with several submodules and routines, plus a detailed online **Help** that can be found at www.ecoevol.ufg.br/sam.wiki (which can be very useful!). The exercises described below are mainly structured according to these modules and submodules. SAM is becoming gradually very large, and we cannot explore it entirely in a single workshop, so we have selected some exercises we considered to be more important and useful (you can explore the entire program by yourself later).

The basic purpose of this tutorial is to guide the students through key exercises in SAM, so that all the basic functions are explored under the supervision of the lecturers. All the exercises here could be repeated using a dataset of your own, but for teaching purposes we will use a standard dataset.

This tutorial assumes that you have successfully downloaded and installed the latest version of SAM from the official website, and have a reasonable knowledge of classical statistical methods (particularly regression). If you have problems or get error messages during any step in the tutorial, please contact one of the instructors to get help.

'DATA' MENU

This includes several sub-modules for a wide variety of data operations, including simple statistics, graphs and maps, transformations, defining connectivity between spatial units, GIS operations and principal components analysis for data reduction.

Building a SAM database

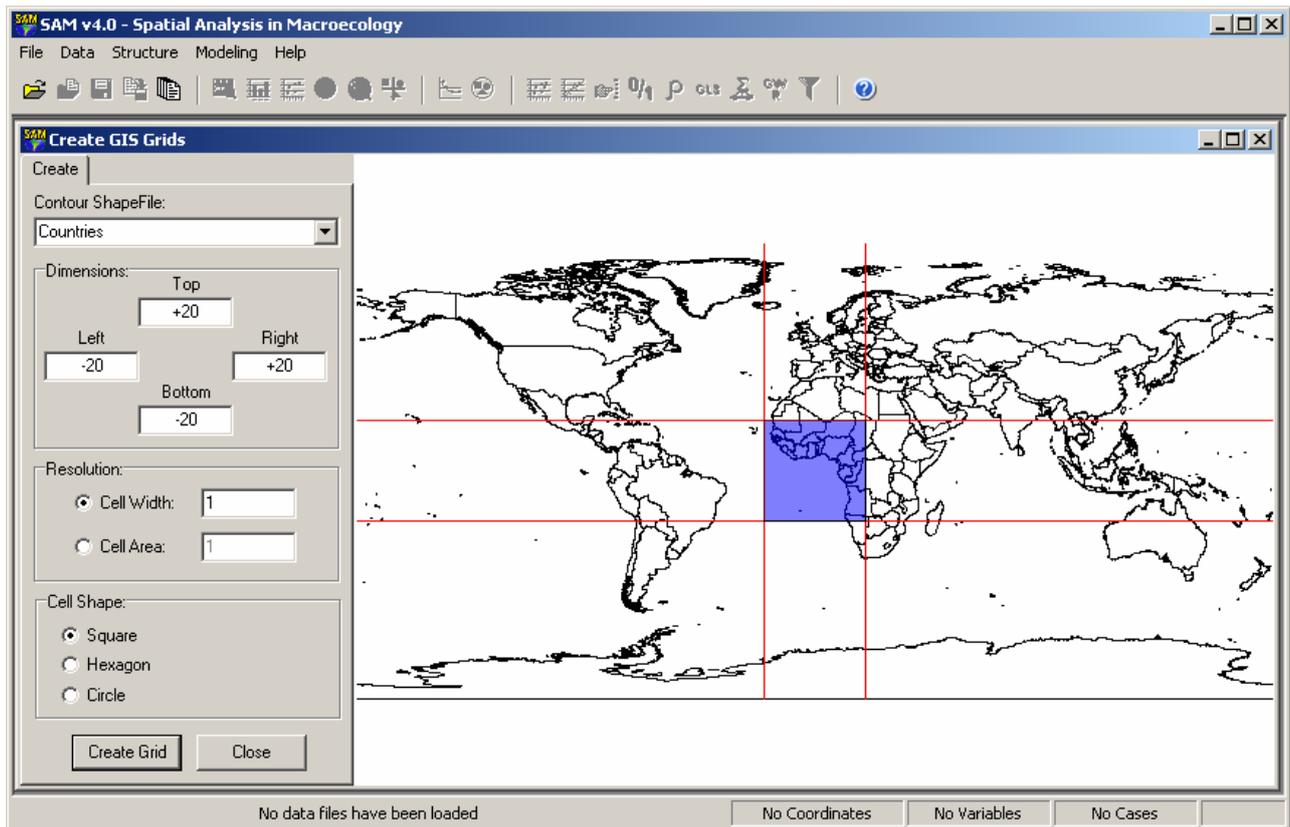
SAM version 4.0 and later are able to build biogeographic databases, using GIS functionality. The necessary data to build the geographic databases are:

- Species geographic ranges: should be formatted in ESRI shapefiles, as polygons or points. For the purpose of this exercise we will use NatureServe data. NatureServe provides free data on the geographic distribution of all Mammals and Birds of the Western Hemisphere, as well as all Amphibians worldwide. You can download NatureServe data from <http://www.natureserve.org/getData/index.jsp>.
- Climatic/environmental data: these data should be formatted in one of the raster grids permissible in SAM: *.bil, *.bip, *.asc, *.xyz, *.txt. For the purposes of this exercise we will use WorldClim data, which provides minimum, mean and maximum monthly temperature and precipitation, altitude and composite bioclimatic variables at various spatial resolutions. You can download WorldClim data from <http://www.worldclim.org/current>

Building a grid

The first task for building your SAM database is to define your grid. A grid is a series of polygons (known as grid cells) that cover your geographic domain of interest. In most cases, grid cells have the same or similar area. In macroecology your domain of interest often has the size of a continent.

In SAM, find the menu Data > Data Handling > GIS Grids > Create GIS Grid. If you have correctly installed SAM in your computer, you should see a window like the one below:



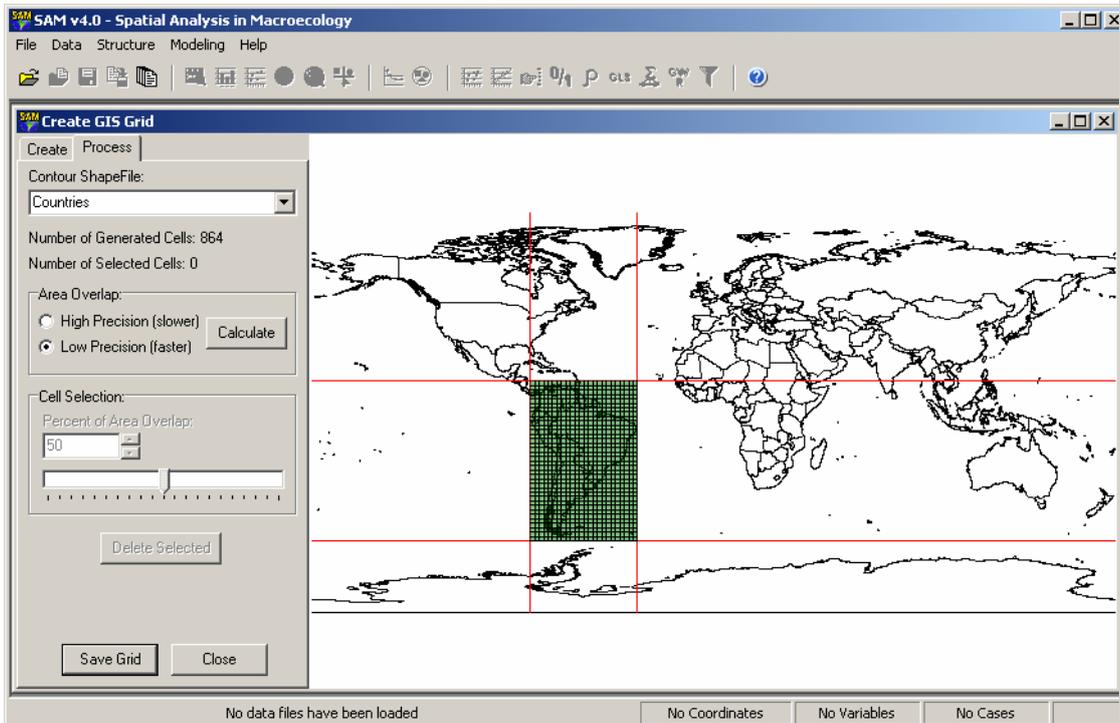
The purple square area in the world map is defined by the coordinate values on the upper right corner of the window (Left: -20, Bottom: -20, Right: +20, Top: +20). The coordinate values are measures in units of degrees of latitude and longitude. This box defines the boundaries of your area of study. Feel free to move this box to South America (or any area within the Western Hemisphere) by changing the values to: Left: -82, Bottom: -58, Right: -34, Top: +14. It is very important that you always set the left value smaller than the right value, and the top value bigger than the bottom value. Notice that by holding <Shift> or <Ctrl> and clicking and dragging the mouse using the left button you can define a square area, and SAM will inform to you the exact coordinates of the box that you have drawn.

For all maps in SAM, you can **use your mouse wheel (or scroll zone on a laptop mouse pad) to zoom in and out in your map**, and you can drag the map using the left mouse button. This should help you to define the limits of your area of interest.

The second step towards defining a grid is cell size. Unfortunately the current version of SAM does not allow building real equal area cells, but only cells with equal latitude and longitude side lengths. For this reason, a cell of the same side length located at 60 degrees of latitude will have approximately half the area of a cell located at the Equator (0 degrees of latitude). We should not

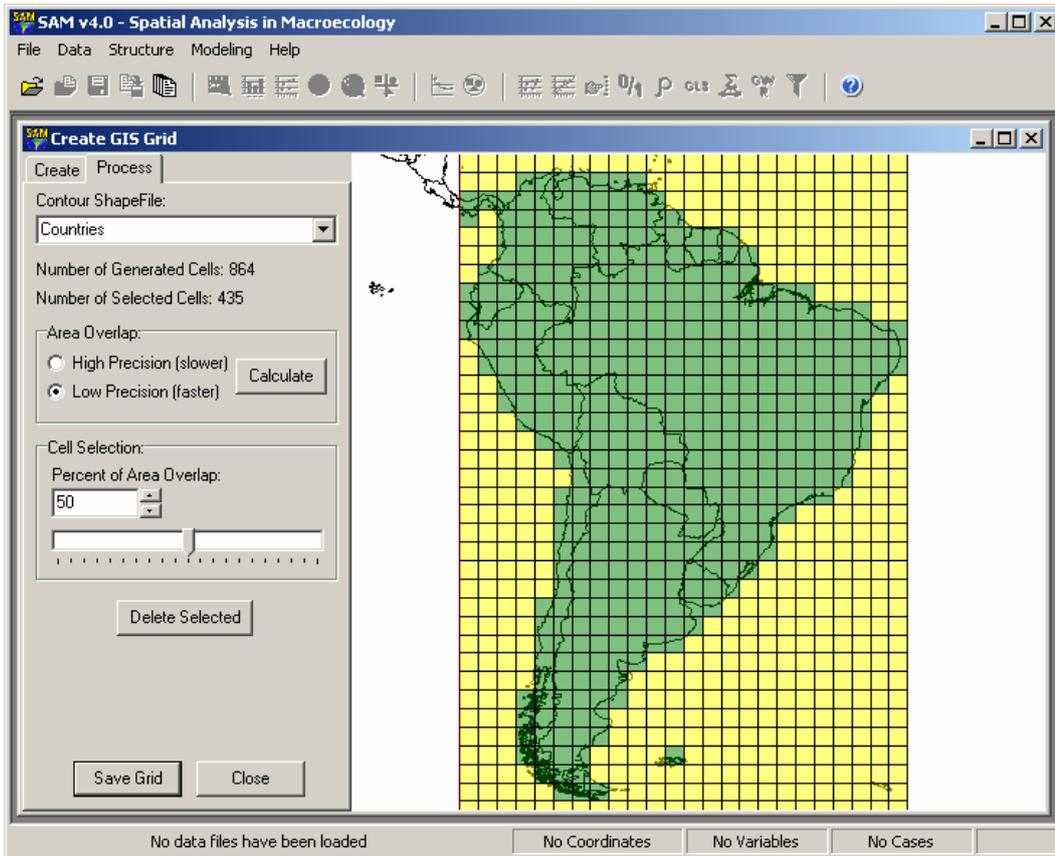
forget this during our analysis, but for the moment we will define the side length of our grid cells as 2 units of degrees of latitude and longitude.

Leave the option “Cell Shape” as “Square” for the moment, but feel free to go back to this option in a later and check what the other options do. As soon as you are ready with the settings, click “Create Grid”. You should now see a window similar to this one:

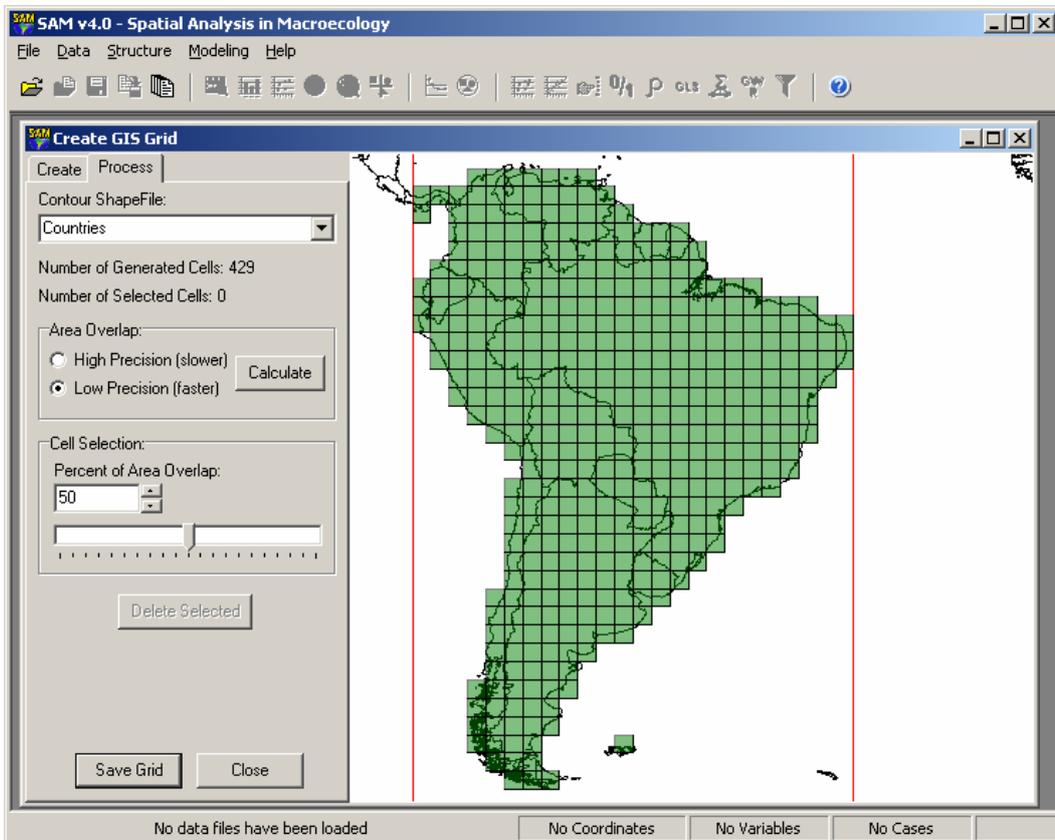


SAM has created a series of small, regular-sized squares on top of South America. From left to right, bottom to top, these grid cells cover the area defined by the bounding box set by you. Feel free to zoom in and out using the mouse wheel, and pan along the continent by clicking the left mouse button and dragging the map. You should notice that not all cells overlap the geographic domain of South America. In fact, some cells are exclusive in the Atlantic or Pacific oceans. We want to get rid of those cells, as they contain no bird or terrestrial mammal species. But how do we do that?

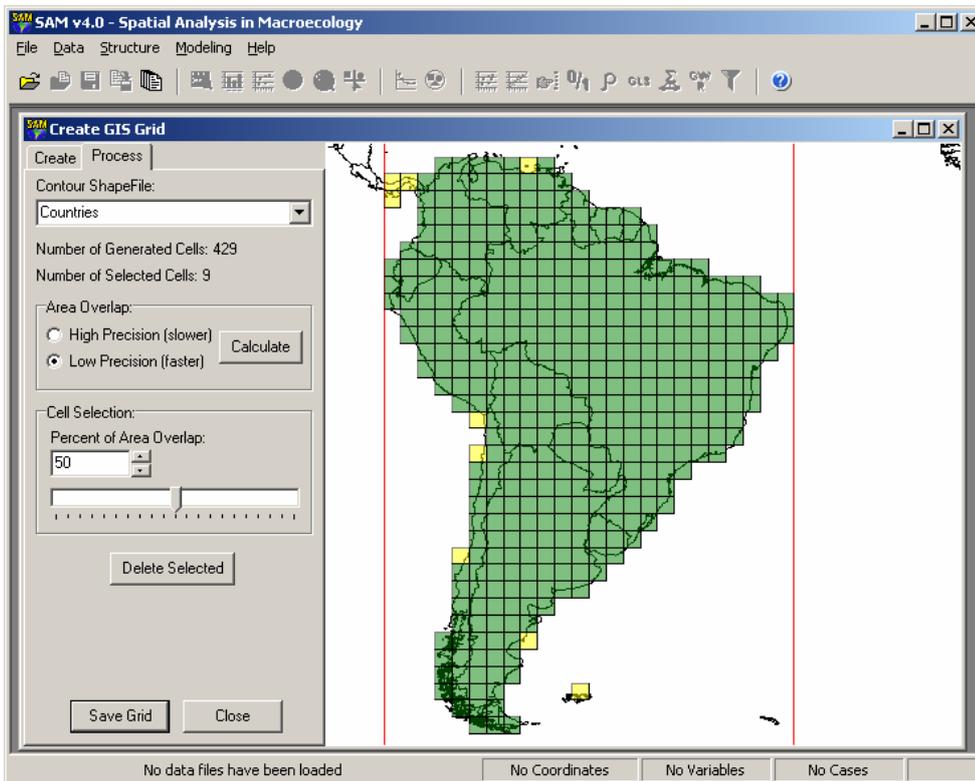
In order to automatically select cells that do not overlap your domain defined by the world shapefile (you can also use your own shapefile, defined in “Contour ShapeFile” box), click the button “Calculate” (leave the option “Low Precision” selected). Those cells that fall outside the South American boundary now look yellow, which means that they are selected:



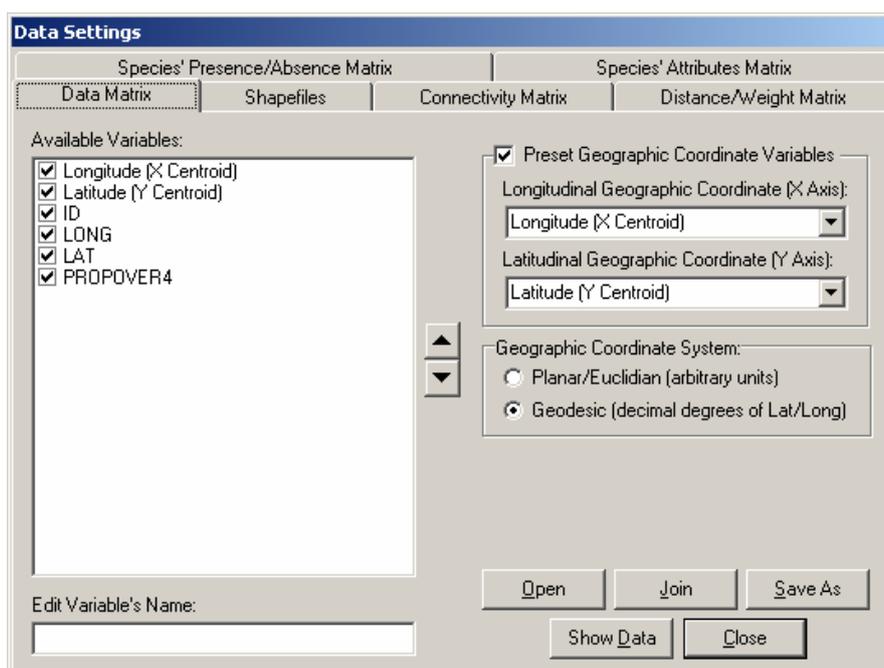
If you click "Delete Selected", those selected cells will be excluded:



Now, hold the *shift* button to select the remaining cells with small land area, such as island and coastal cells. I will also select three species that overlap Panama:



When you click the button “Save Grid” a file save dialog pops up. Define the location and name of the shapefile that describes your grid, and save it for later analysis. After saving the file, the “Data Settings” window will pop up, showing the variables available in the shapefile. Notice “Longitude (X Centroid)” and “Latitude (Y Centroid)” selected as “Preset Geographic Coordinate Variables” and “Geodesic” as the default “Geographic Coordinate System”.



Rescaling environmental data

Once your grid system is defined, the next step to build a SAM database is to get environmental data in the same scale and position as your grid is defined. In other words, the goal is to get an environmental descriptor of each map cell. However, most of the environmental data available on the internet is at a much finer spatial resolution. So we will have to get an averaged value for each environmental variable within our 2 x 2 degrees grid cells. Of course, a descriptor of environmental variation within the cell will also be welcome. The environmental data should be in any of the raster grids allowed in SAM, such as the standard *.bil and *.bip file formats. Those formats are available for download in the online WorldClim database (worldclim.org/download). I have downloaded the “generic grids” in “10 arc-minutes resolution”, which is enough for the resolution of our macroecological grid. If you have not done so already, download the “Bioclim” one at this resolution and unzip the file.

The SAM module designed to process raster grids is located at Data > Data Handling > GIS Grids > ReScale Raster Grid.

On the ReScale Raster Grid window you will define what your grid is. As you recall, your grid is composed of polygons that define an area within your domain of study. One could use simple latitude and longitude coordinates of the centroid of each of those areas, but that would only work if your grid system is regular. However, because we have created a grid and saved it as a shapefile, we have not only the coordinates of each cell centroid, but the coordinates of the entire polygon (actually, the four vertices of each polygon). Thus, in our ReScale Raster Grid module, select “Use Shapefile” and then select the shapefile that you have created for your grid.

The next step is to select your environmental file. As mentioned before, WorldClim provides data in the standard *.bil raster format. Click on the upper-right corner button  and navigate to the bioclim files you downloaded from the WorldClim website. You can select all using <Shift>, in the normal way, and open them all together. Here is what each variable means:

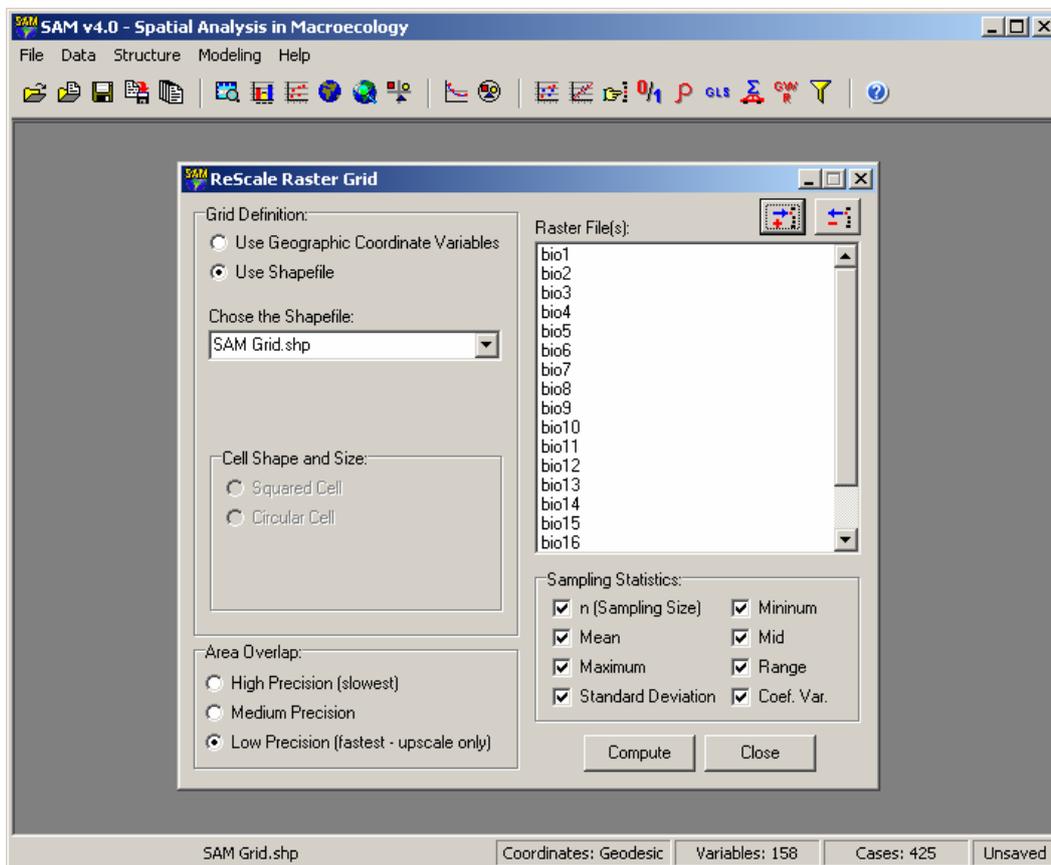
BIO1 = Annual Mean Temperature

BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

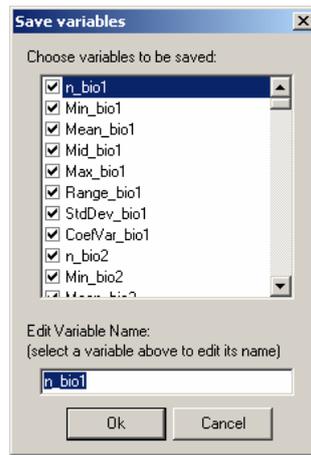
BIO3 = Isothermality (P2/P7) (* 100)

BIO4 = Temperature Seasonality (standard deviation *100)

- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (P5-P6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO10 = Mean Temperature of Warmest Quarter
- BIO11 = Mean Temperature of Coldest Quarter
- BIO12 = Annual Precipitation
- BIO13 = Precipitation of Wettest Month
- BIO14 = Precipitation of Driest Month
- BIO15 = Precipitation Seasonality (Coefficient of Variation)
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

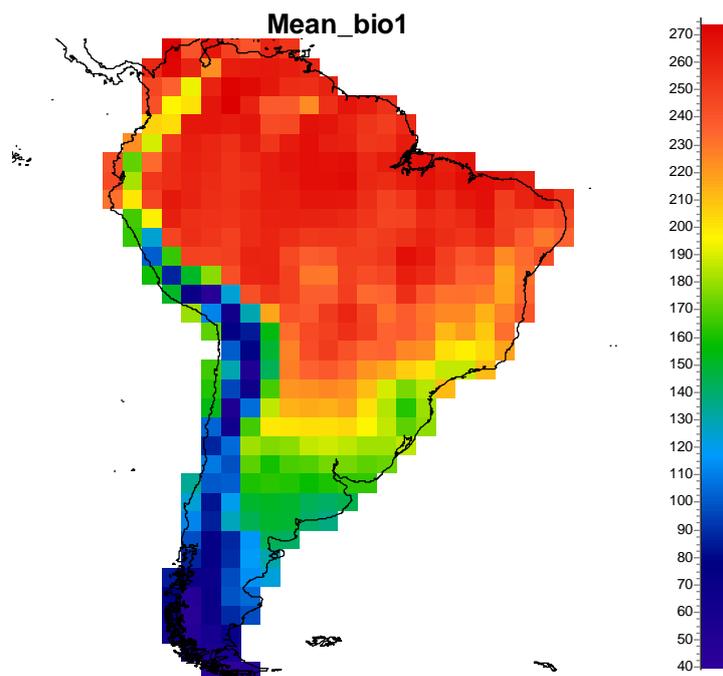


Because SAM knows exactly what the boundaries of each grid cell are, it is able to select all raster points that fall within each grid cell. After finding those raster points in the original bioclimatic variables, SAM will get descriptive statistics from the selected points, based on your choice on the lower-right corner ("Sampling Statistics"). Click "Compute" and wait a few seconds to get this:



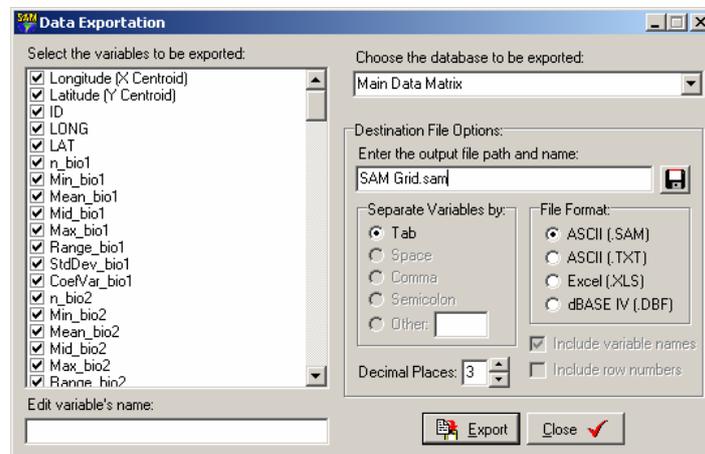
One hundred and fifty two new variables have been created! Thus, for each cell in our grid, we have eight new descriptors. Those represent, in the same order as displayed above, (1) number of raster points found within each grid cell, (2) the minimum value, (3) the mean value, (4) the median ('mid') value, (5) the maximum value, (6) the range in values, (7) the standard deviation and (8) coefficient of variation. Feel free to uncheck each variable, and rename those that you think are interesting. Click "Ok" and close the ReScale Raster Grid window.

Go to Data > Graphs and Maps > Map Data Matrix. The first two boxes on the upper-left define the coordinate variables of each grid cell. The lower box, on the lower-left, defines the attribute of the cells to be mapped. Select "Mean_bio1" (which is mean annual temperature). This map should look like this:



Notice that the scale is not in units of degrees Celcius, but degrees Celcius multiplied by 10. Take some time to explore the other variables, rename (module Data > Data Settings) them, and discard those that you think are not interesting.

After doing all this work with environmental variables, let us save all the data to the hard drive! Go to “File > Save As” and in File Format select “ASCII (.SAM)”. Choose the same folder in your hard drive where you have saved the shapefile, and name your *.SAM file using the same you gave to the *.SHP file. SAM is programmed to ignore the *.DBF file associated with the *.SHP file if it finds a *.SAM file of the same name in the same folder. Therefore, the *.SAM file will now become the main database associated with your grid shapefile.

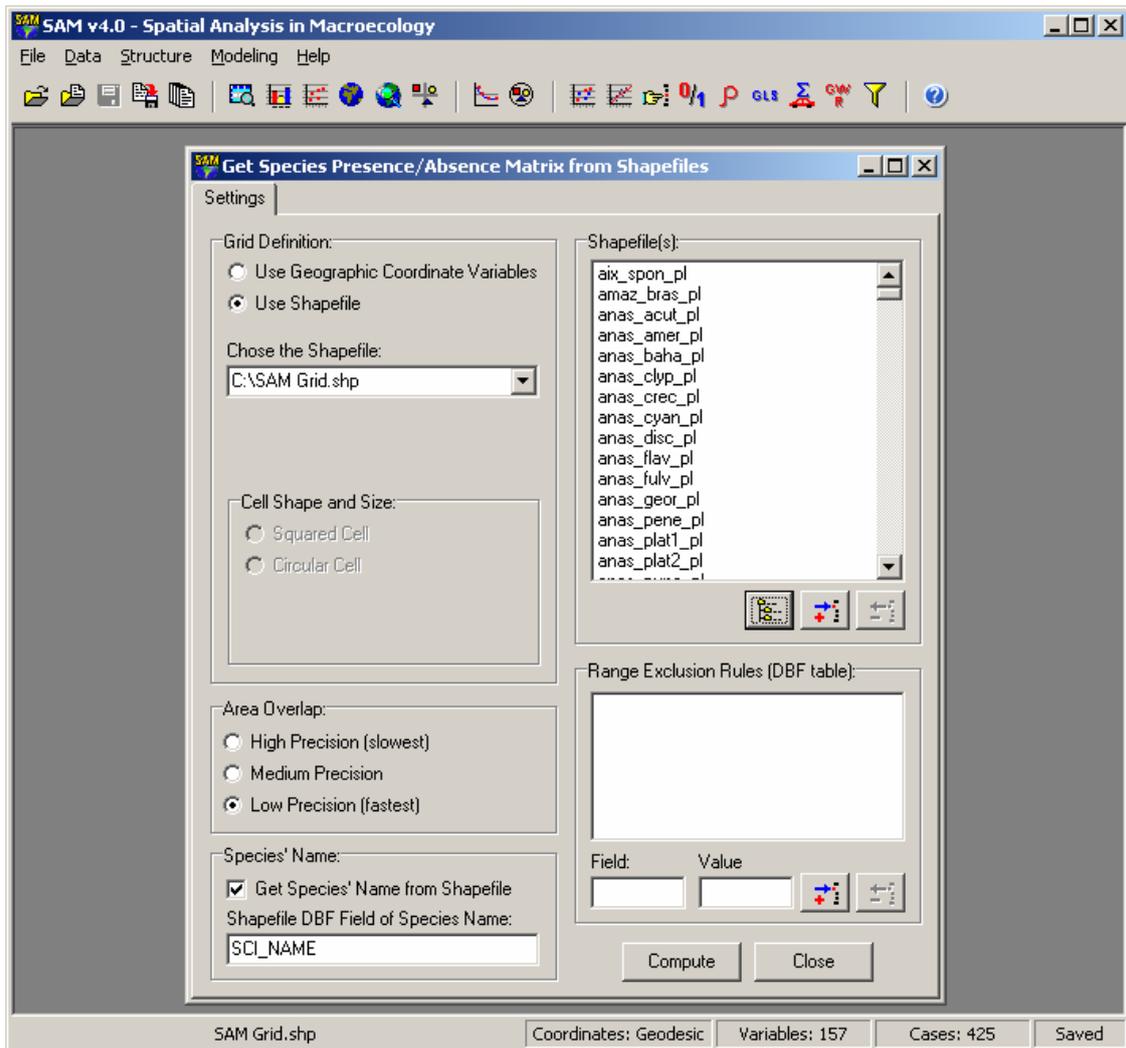


Clipping and “rasterizing” species distributions to your grid

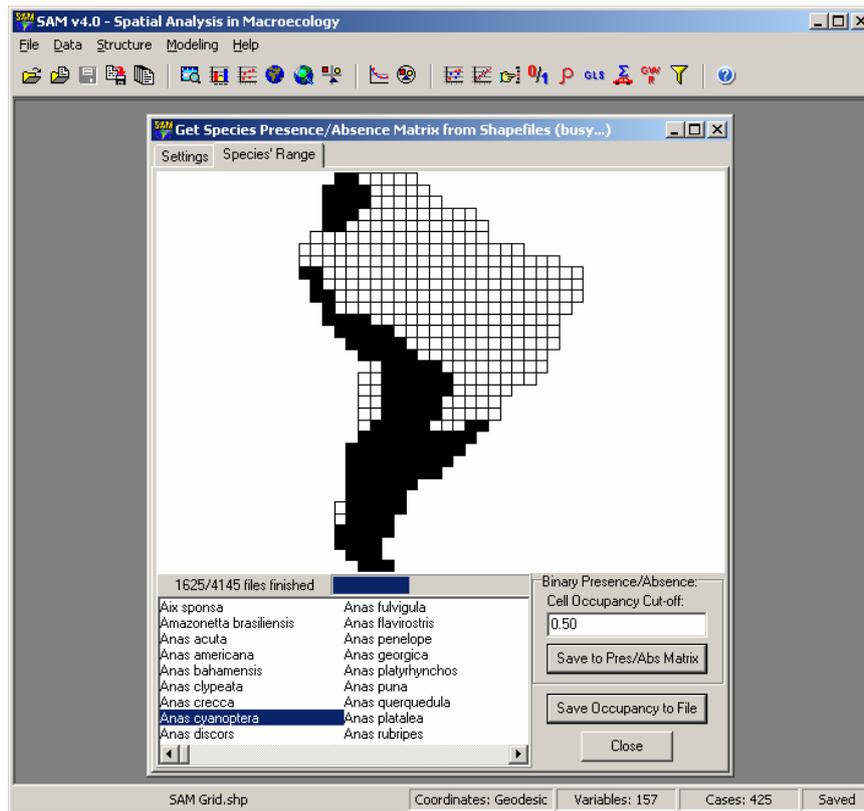
The next step is to get your response variable, or the variable that you will be investigating. Of course, the choice of this variable depends on your scientific question, scale of analysis and expertise. We can use the NatureServe (www.natureserve.org/getData/birdMaps.jsp) database on the geographic distribution of birds in the Western Hemisphere in order to calculate the spatial pattern in species richness of all birds in South America. It is easiest to download the entire database and unzip all the files into a single folder.

In the module “Data > Data Handling > GIS Grids > Get Species Occurrence from Shapefile”, select the shapefile of our newly created grid under “Grid Definition”. Also, under “Species’ Name”

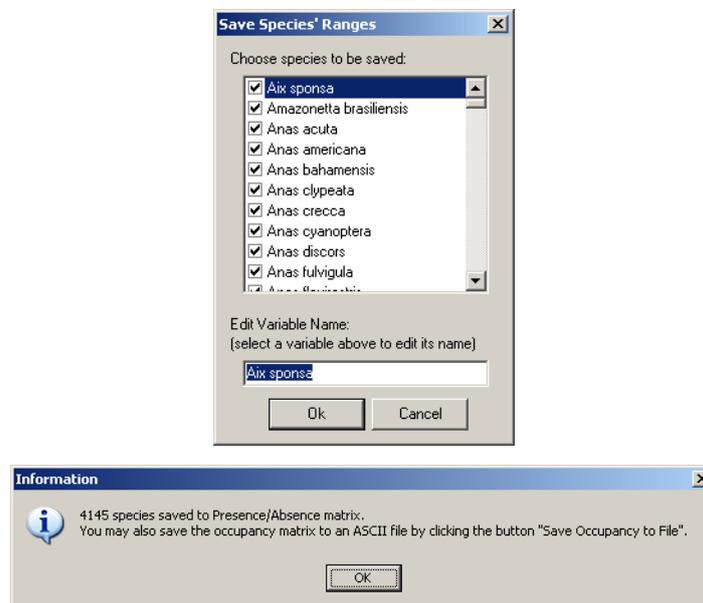
select “Get Species’ Name from Shapefile” and leave “SCI_NAME” as the “Shapefile DBF Field of Species Name”.



SAM will start working on the 4145 files as soon as you click “Compute”. On each of our laptops it took around 10 minutes to process all the files. During the time of processing you can look at the geographic distribution of any species already done, and available in the list. When you click on a species, SAM will show the geographic distribution of that species. Here, for example, is the map for “*Anas cyanoptera*”:



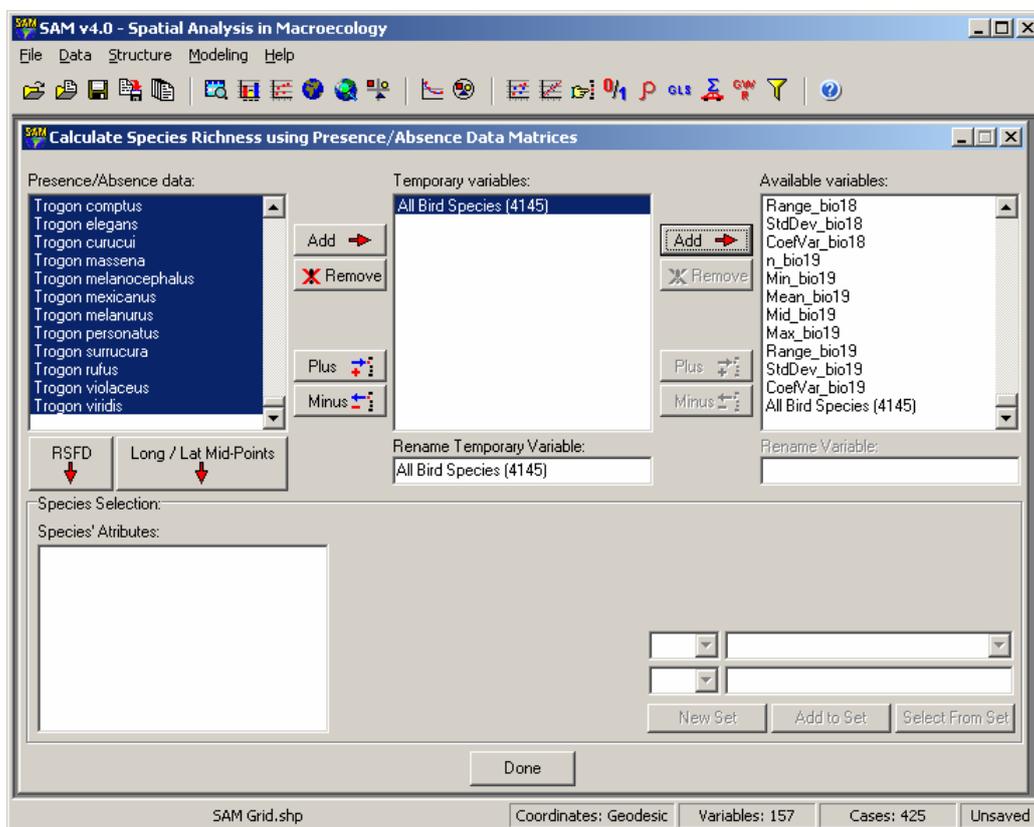
Once SAM finishes processing all files, you should get a window named “Save Species’ Ranges”. This shows the names of all species that have been found in all shapefiles selected for processing.



After clicking “Ok” we see that 4145 species have been saved as a Presence/Absence matrix. This means that the data are not directly available for mapping in the “Main Data Matrix”, but are saved in a separated part of the memory. If you look in “File > Data Settings”, under the “Species’ Presence/Absence Matrix”, you should find all species that were processed from the shapefiles.

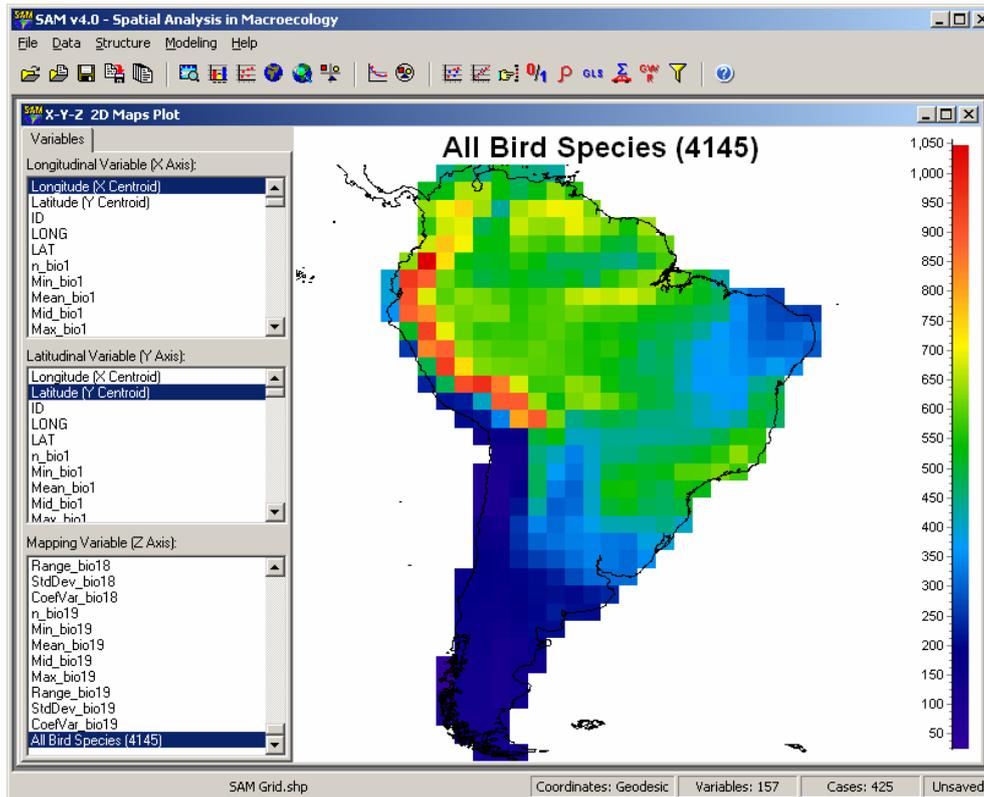
Notice that you can also save the species' Presence/Absence Matrix to a ASCII file, by clicking in the "Save Occupancy to File" button. You will be prompted a list of species and then a file save dialog, in which you should choose the name and path of a text file that will hold the binary matrix with zeros for absence and ones for presence.

But how can one calculate species richness from these data? In the module Data > Data Handling > Species Occurrence > Calculate Species Richness you can sum across all columns of the Presence/Absence Matrix that is stored in SAM's memory:



Select all or some of the species available in the upper-left corner, where all species available are listed. Once you select the species, click the left "Add" button and a new temporary variable appears in the upper-mid box, named. This temporary variable already holds the species richness of those species selected before. However, sometimes you want to subtract or add species richness from a second pool of selected species, to get the richness of a sub-group, and to do so you can use the "Plus" and "Minus" buttons. When you are ready, select the variable in the middle box and click the right "Add" button. The variable should now be listed with the other variables already available.

As you may already have guessed, the newly created variable (bird species richness) can now be mapped. Go to “Data > Graphs and Maps > Map Data Matrix” and select the new variable at the bottom of the lower-left box. You should get your species richness map. Of course, feel free to explore all graphing options available in SAM.



Dealing with files and file formats

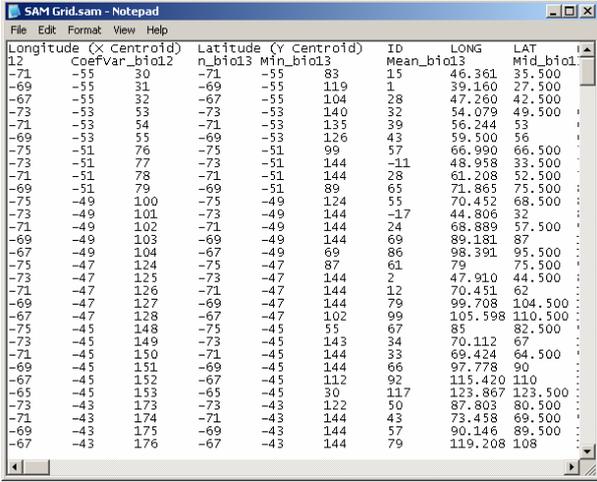
Now that you have produced your own database, maybe we should consider data formatting. SAM input files can be formatted as ASCII (American Standard Code for Information Interchange), DBF (Database File), EXCEL or ESRI shapefile. As in most statistical software, variables are arranged in columns, and cases (spatial units) in rows. The first row contains the name of each variable in the file. An important issue regarding input data is that, because SAM was developed for spatial analyses, your data file must contain two geographic coordinate variables in order to compute spatial statistics. If these two coordinates are named “Latitude” and “Longitude”, then SAM will understand that these variables are expressed as degrees (and decimal degrees) of latitude and longitude (for example, 10.5 indicates 10°30′). A positive latitudinal value means a site located north of the Equator, and a negative south; positive and negative values for longitude mean, respectively, east and west of the Greenwich meridian. In SAM, when coordinate variables are named “Latitude” and “Longitude”, all distances calculated within SAM will be geodesic distances, which take into account the Earth’s surface curvature. It is highly recommended to use this option whenever possible, especially in studies at large spatial scales. If you are using a planar system (i.e. Euclidean, arbitrary scale), you can use other names (e.g., “X Axis” and “Y Axis”, or “Vert Axis” and “Horiz Axis”) to express arbitrary spatial coordinates and, in this case, SAM will compute geographic distances using a planar Euclidean system (the distance between two sites is the length of a straight line that links them). Of course, you can easily change this setting, at any time in SAM (Menu **File** > **Data Settings**); whatever you define will be used as default in all SAM routines thereafter.

Today, most macroecologists use ESRI shapefiles to define grids and store biogeographic and environmental data, which is exactly what we did when we build our database. As mentioned before, ESRI shapefiles are composed of at least three files: *.shp, *.shx and *.dbf. The first two are designed to store the geographic information of the vertices of polygons or points, whereas the latter stores the data matrix associated with the geographic objects. In SAM, only the numeric information contained in the DBF matrix is used. Actually, DBF matrices are quite old and outdated today, and impose several constraints on how the numbers are formatted. For this reason, whenever there is a xxx.shp file along with a xxx.sam file (xxx is any name of file), SAM will actually use the xxx.sam file as input to the numeric database, ignoring the *.DBF file. This provides flexibility to the size of the database that can be used, as well as name of variables.

Computational capacity: some of the SAM routines are computationally complex, and will be time consuming when used for large datasets (especially when $n > 2000$ spatial units), depending of course on your computer (the main limitation is actually the available RAM memory, and computer time will depend on processing speed [GHz]). In our experience we can deal with matrices with more than 5000 spatial units using good computers (RAM > 3 Gb), although some analyses, such as autoregressive models and eigenvector spatial filtering, can last longer than 4 hours in some cases! Thus, for classroom exercises, we recommend data with no more than 500 spatial units, to allow an effective and feasible exploration of SAM modules and routines.

Getting familiar with files in SAM

By using a text editor (such as Notepad or Microsoft Word) open the dataset file in SAM format that you created in the exercise above. I have named my file as "SAM Grid.SAM". Now, using Microsoft Excel, open that same file. You should see that a SAM file is nothing more than a text (ASCII) file that have columns separated by tabs, while rows are in different lines. Also, in Excel you should see that the entire dataset is in the upper left corner of the spreadsheet. **No cells within the data frame can contain missing values, and all cells except the column headings must only contain numeric data.** The only difference between a *.SAM file and a *.TXT file is that the former can be double clicked and opened in SAM.

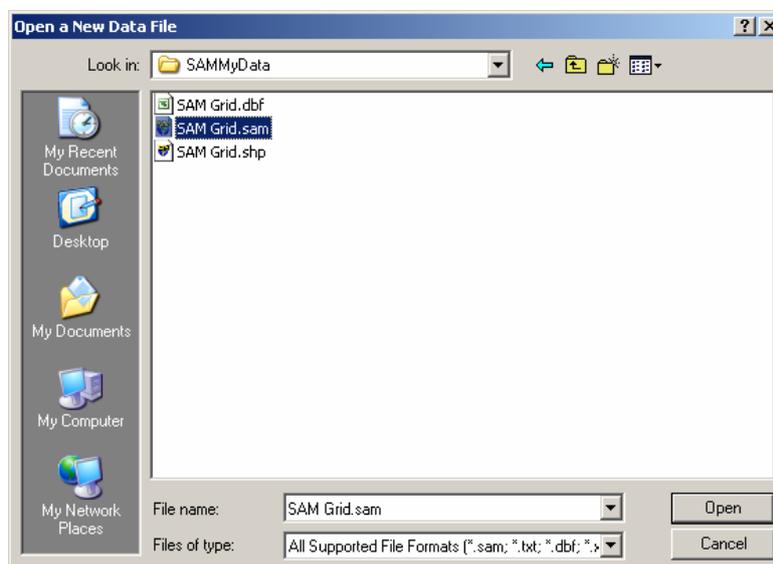


Longitude (X Centroid)	Latitude (Y Centroid)	ID	LONG	LAT
12	30	15	46.361	35.500
-71	-55	83	1	39.160
-67	-55	119	28	47.260
-73	-53	104	32	54.079
-71	-53	140	39	56.244
-71	-53	135	43	59.500
-69	-53	126	57	66.990
-75	-51	99	-11	48.958
-73	-51	144	28	61.208
-71	-51	144	65	71.865
-69	-51	89	55	70.452
-75	-49	124	-17	44.806
-71	-49	144	24	68.889
-69	-49	144	69	89.181
-67	-49	104	86	98.391
-75	-47	124	61	79.750
-73	-47	144	2	47.910
-71	-47	144	12	70.451
-69	-47	144	79	99.708
-67	-47	102	99	105.598
-75	-45	148	85	82.500
-73	-45	143	34	70.112
-71	-45	144	33	69.424
-69	-45	144	66	97.778
-67	-45	112	92	115.420
-65	-45	133	117	123.867
-73	-43	173	50	87.803
-71	-43	174	43	73.458
-69	-43	175	57	90.146
-67	-43	176	79	119.208

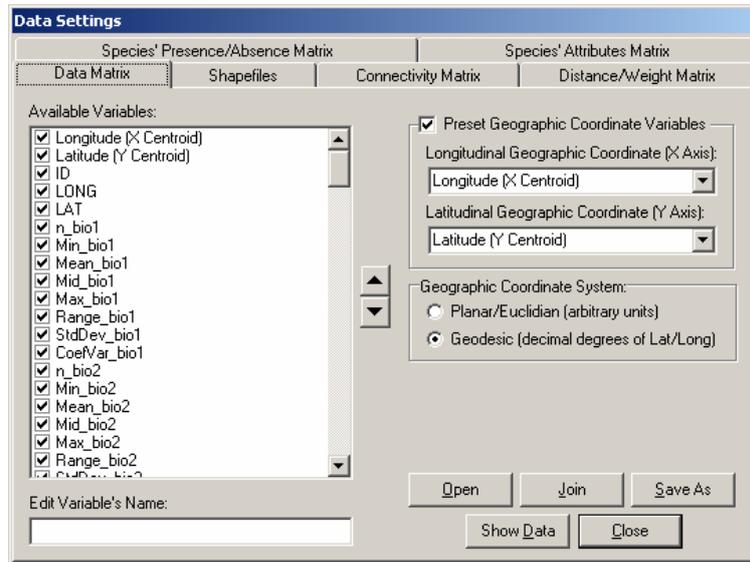
	A	B	C	D	E	F	G	H	I
1	Longitude	Latitude (YID)	LONG	LAT	n_bio1	Min_bio1	Mean_bio1	Mid_bio1	
2	-71	-55	30	-71	-55	83	15	46.361	35.5
3	-69	-55	31	-69	-55	119	1	39.16	27.5
4	-67	-55	32	-67	-55	104	28	47.26	42.5
5	-73	-53	53	-73	-53	140	32	54.079	49.5
6	-71	-53	54	-71	-53	135	39	56.244	53
7	-69	-53	55	-69	-53	126	43	59.5	56
8	-75	-51	76	-75	-51	99	57	66.99	66.5
9	-73	-51	77	-73	-51	144	-11	48.958	33.5
10	-71	-51	78	-71	-51	144	28	61.208	52.5
11	-69	-51	79	-69	-51	89	65	71.865	75.5
12	-75	-49	100	-75	-49	124	55	70.452	68.5
13	-73	-49	101	-73	-49	144	-17	44.806	32
14	-71	-49	102	-71	-49	144	24	68.889	57.5
15	-69	-49	103	-69	-49	144	69	89.181	87
16	-67	-49	104	-67	-49	69	86	98.391	95.5
17	-75	-47	124	-75	-47	87	61	79	75.5
18	-73	-47	125	-73	-47	144	2	47.91	44.5
19	-71	-47	126	-71	-47	144	12	70.451	62
20	-69	-47	127	-69	-47	144	79	99.708	104.5
21	-67	-47	128	-67	-47	102	99	105.598	110.5
22	-75	-45	148	-75	-45	55	67	85	82.5
23	-73	-45	149	-73	-45	143	34	70.112	67
24	-71	-45	150	-71	-45	144	33	69.424	64.5
25	-69	-45	151	-69	-45	144	86	97.778	90
26	-67	-45	152	-67	-45	112	92	115.42	110
27	-65	-45	153	-65	-45	30	117	123.867	123.5
28	-73	-43	173	-73	-43	122	50	87.803	80.5
29	-71	-43	174	-71	-43	144	43	73.458	69.5
30	-69	-43	175	-69	-43	144	57	90.146	89.5
31	-67	-43	176	-67	-43	144	79	119.208	108
32	-65	-43	177	-65	-43	104	120	131.442	130.5
33	-73	-41	197	-73	-41	140	52	98.914	88

You should always remember that SAM is distributed with Sample Datasets, which can be used as a model for formatting your own datasets. Please refer to those sample datasets if you need to check rules for file formatting.

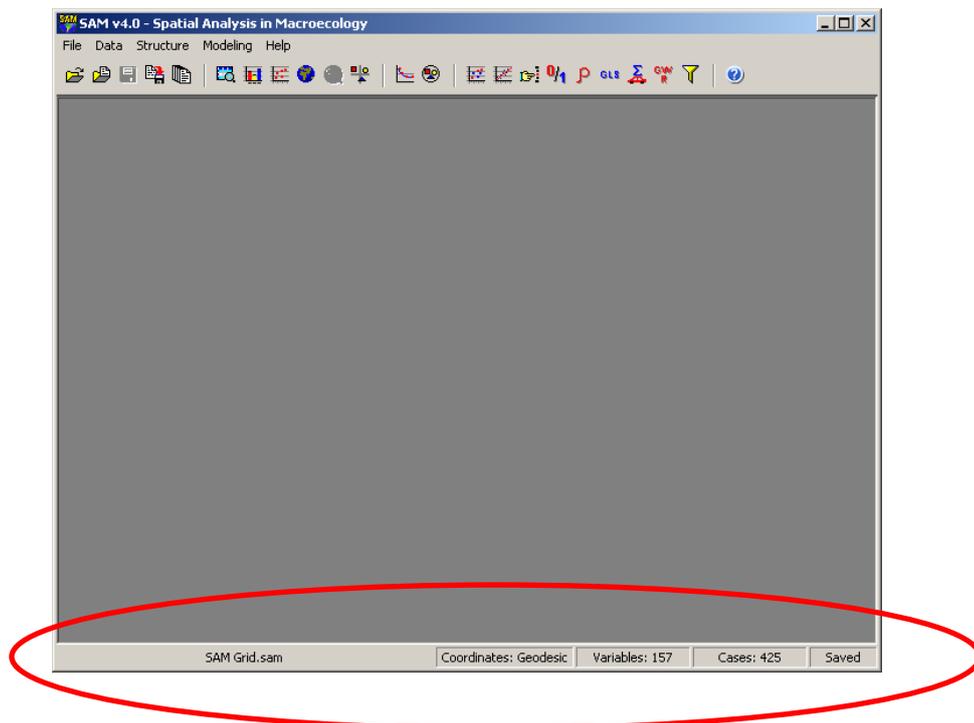
Now that you know how files should be formatted for SAM it is time to see how to open them. There are several ways you can open a dataset in SAM. You can use the menu **File > Open > Open a New Main Data File** (or simply double-click the dark-grey area in SAM). Go ahead and open the file "SAM Grid.SAM" in SAM and notice how the information appears in the bar at the bottom of the screen (filename, type of spatial coordinates, number of variables and cases).



The first thing you get when you open a dataset (successfully) is the “Data Setting” window, which is useful to confirm the variables available for analysis in SAM.



This is a very important window, with which you should become familiar. In the left-hand side you have the list of variables available for analysis, and each can be un-ticked if you do not want it in your data. In the top of the right-hand side you can choose the geographic coordinate variables, and preset them for all the analysis in SAM. By default, SAM will search for variables named “Latitude” and “Longitude”, or variables starting with “X” and “Y”. In our case, SAM identified the variables “Longitude (X Centroid)” and “Latitude (Y Centroid)”, and guessed that these are in a “Geodesic” coordinate system.



In the status bar at the bottom of the screen, you can check the filename, the geographic coordinate system in use (geodesic, in this case), the number of variables and cases. SAM also shows that the data currently uploaded are the same as those found in the file (it says “saved” in the lower-right corner), and as soon as you add or change the data it will change to “unsaved”.

Files can be saved at anytime, either overwriting the current file, or saved as a new file, with a new file name. The “File > Save As” window allows selecting the file name and path, file format, as well as the variables that are to be stored in the new file.

Managing and exploring data

Menu: Data > Data Handling

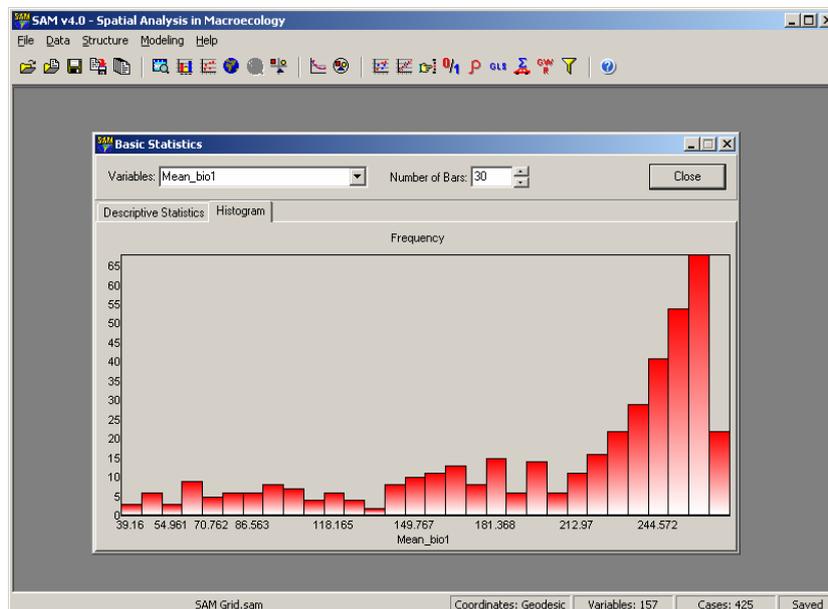
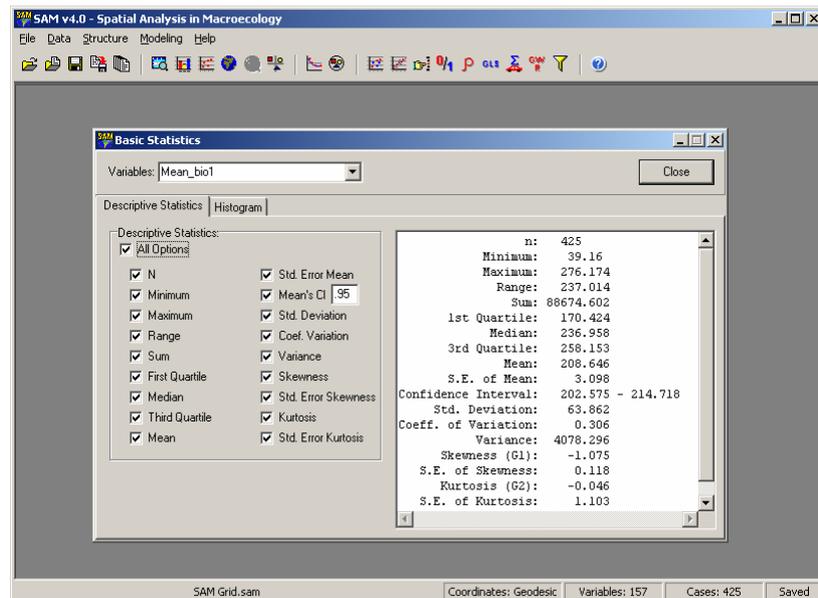
There are many different modules available under the menu Data > Data Handling. In general, those modules concern transforming, producing and modifying new variables. Below is a brief explanation of what most of them do. Of course, feel free to explore them in more detail.

- a) *Show Data Matrix*: Shows the current Data Matrix in a spreadsheet. You can check and change data, but you should probably use Excel if you are planning a systematic change of your data matrix.
- b) *Transformations*: Simple mathematical transformations that can be very handy when you need to fit a particular distribution. On the left side of the window, select among one of the available mathematical functions. On the right side, select which variable(s) you would like to transform. In the lower-right corner, select whether you want to transform the variable itself, overwriting it, or create a new variable, leaving the original intact.
- c) *Polynomial Expansions*: This module can be very useful if you want to create polynomial expansions of one or more variables, particularly for geographic coordinate variables. These polynomial expansions can be used in Trend Surface Analysis, for example. To expand variable(s), select one or more variables, select the polynomial order (2 or greater), and click “Generate”. The generated variables will appear in the right box. Among the “Temporary Variables”, select those that you wish to use, and click “Save”.

- d) *Distance / Similarity Matrices*: In this module you can calculate distance or similarity between pairs of sampling sites, using one of the dozens of available distance and similarity metrics (e.g. Euclidean, Geodesic, Nei distance indices, or Jaccard's, Sorensen's, Faith's, Simpson's similarity indices). This module is particularly important for generating distance/similarity matrices to be used in Mantel Tests. Select the variables that describe your sampling units, choose among "distance metrics" and "similarity coefficients", choose the particular function you want to use in the option box, and click "Calculate". You have the option of renaming the distance/similarity matrix just created by selecting it and changing its name in the "Matrix Name" box.
- e) *Geographic Connectivity/Distance Matrix*: This module is used to Create, Edit and Plot Connectivity and Distance Matrices (studied in more detail below).
- f) *Species Occurrence*: The sub-module "Calculate Species Richness" can be used when a "Species Presence/Absence Matrix" is loaded into SAM (File > Data Settings; see above), which consists in a binary matrix in which species are columns and sampling units are rows. Using these datasets you can calculate species richness patterns by selecting species individually, based on some attribute of the species, or selecting all species. The sub-module "Map Species' Attributes" can be used to plot on the map the mean or variance of a particular attribute, such as body size, of all species that are present in a given area (a species' assemblage or community). For this module you need to load not only the "Species Presence/Absence Matrix", but also the "Species Attributes Matrix" (see File > Data Settings).
- g) *GIS Grids*: Can be used to create a GIS grid and to resample environmental data into this new grid, as well as to get species occurrence from shapefiles (see above for more).

Basic Descriptive Statistics

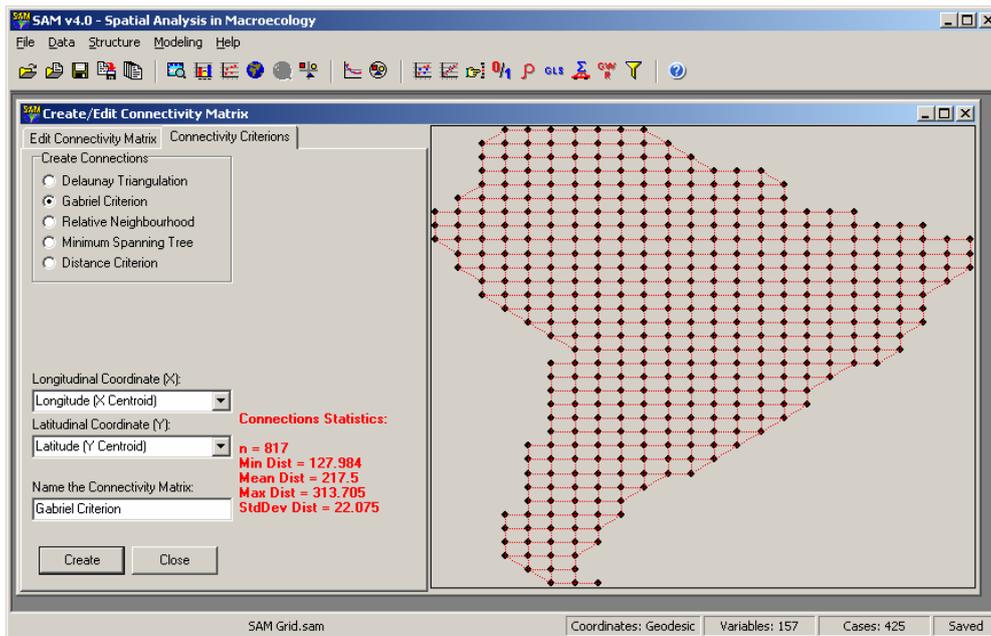
A simple but useful module is Data > Descriptive Statistics. In this module you can find descriptive analytical statistics and a configurable histogram:



The example above shows the descriptive statistics (average, min, max, standard deviation, variance) for Mean Annual Temperature in South America, as well as the distribution of values across map cells.

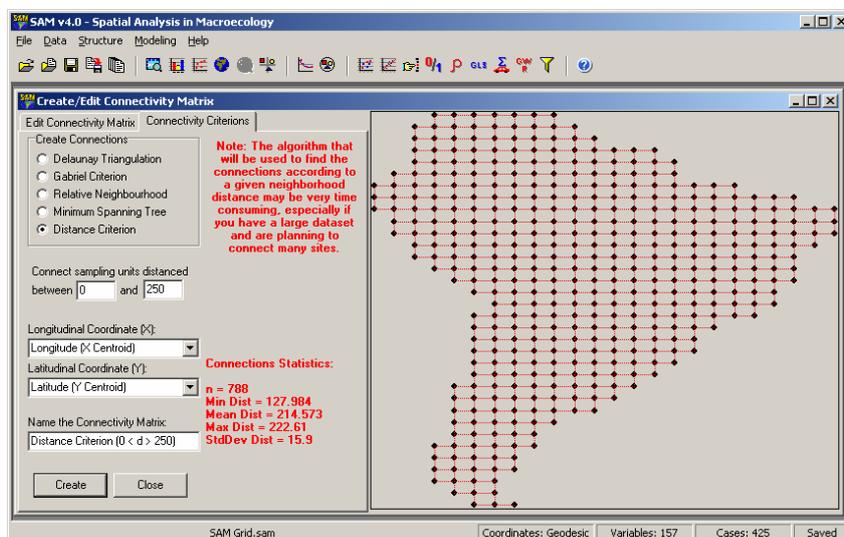
Designing a connectivity matrix

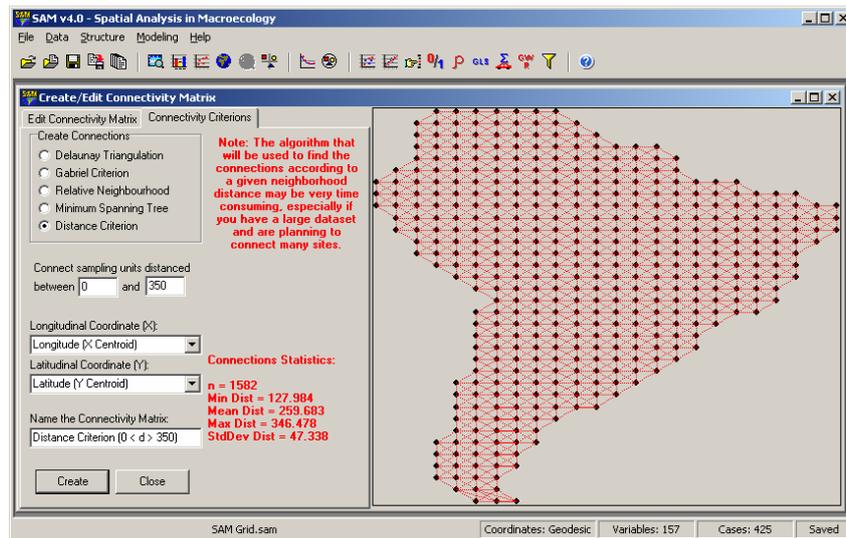
More importantly, in the menu **Data > Data Handling > Connectivity/Distance Matrix > Create/Edit Connectivity Matrix**, you can explore different possibilities for creating and editing connections between cells, using different criteria. Click on the “Connectivity Criteria” tab, and select “Gabriel Criterion” (this is one of the faster to calculate, but it will still take a while).



The Gabriel connectivity is displayed, as well as statistics for the connection. Use the button “Create” and these connections will be available to be used by other modules of SAM. Remember that the text present in the “Name of the Connectivity Matrix” text box will become the name of your connection scheme. Use this module and click in the ‘Distance Criterion’, and set the distance limits to 0 and 250 (the units are km), and see the connection scheme changing in the graph on the right. A connectivity criterion, as here, that connects adjacent cells in the vertical and horizontal direction is known as “Rook” (as in the chess game), while connections of adjacent cells in the horizontal, vertical and diagonal are known as “Queen” (in this case change the distance to 0 and 340). Save the “Queen” connectivity criterion so that it can also be used later.

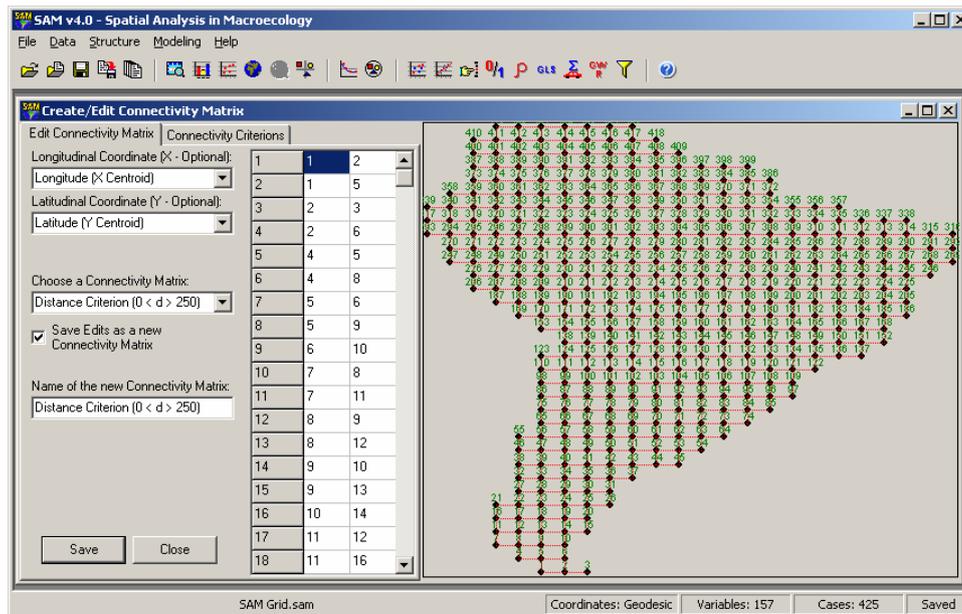
Make sure you click “Create” button to save the connectivity criterion for later analysis.





Another handy tool is to use the mouse to zoom in the map to take a look in details of the connectivity structure. You can activate zoom in the figure by using the mouse wheel or the scroll up/down portion of the mouse pad on most laptop computers: scrolling up (with the cursor over the map) will zoom in; scrolling down will zoom out. You can drag the map by clicking and holding the left mouse button. Also try right-clicking on the map.

Finally, you should know that you can enter your connectivity matrix by typing the connections manually, or change an existing connectivity scheme by manually changing which cells that are connected. This can be done in the “Edit Connectivity Matrix” tab, and can be useful when you have a specific hypothesis for the geographical relationship among your sampling sites (the Matrix W). Explore it as you wish. You can start by either typing numbers in the columns (a lengthy process!) or by changing the numbers of an already existing connectivity matrix (“Choose a Connectivity Matrix” box). Make sure you click “Save” when you are done.

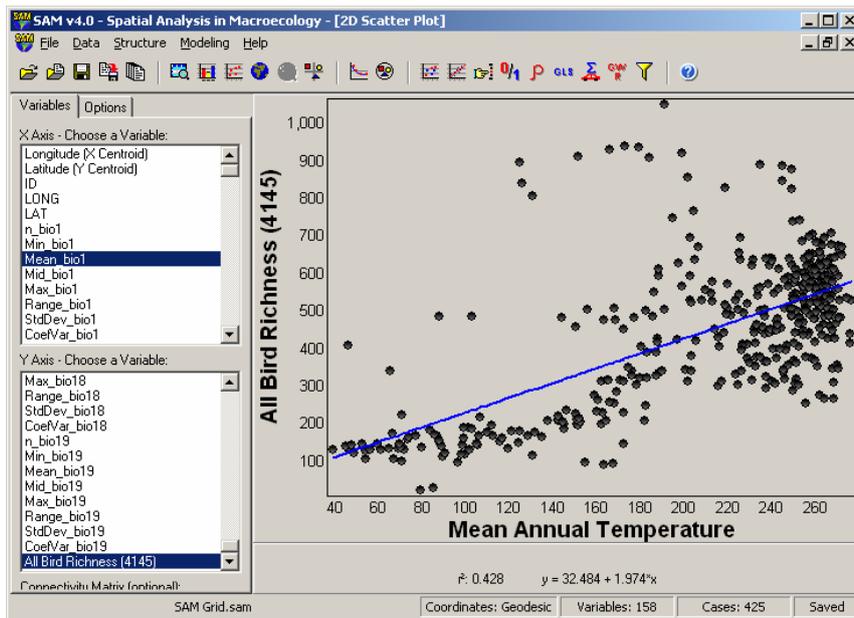


Exploring graphical features in SAM

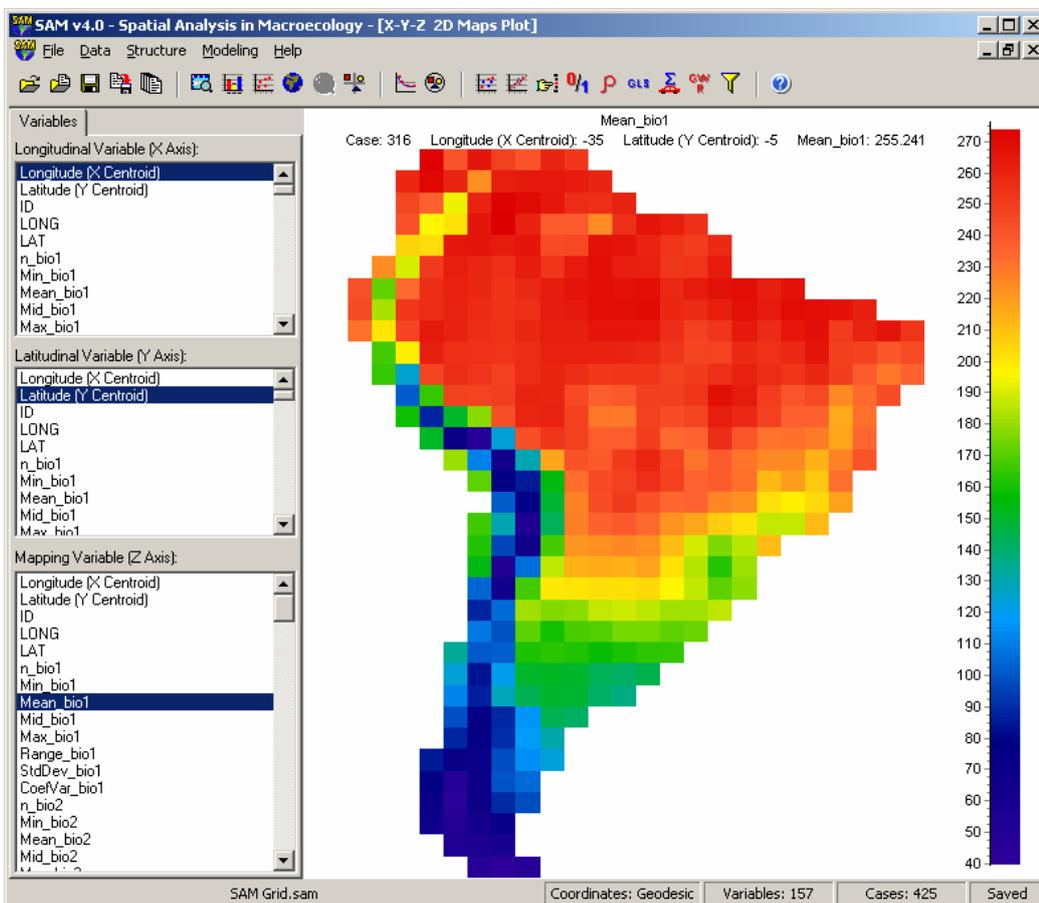
Among the most-used modules in SAM are those located at “Data > Graphs and Maps”, where you can build maps (“Map Data Matrix”), 2D and 3D scatterplots, create bar plots (where individual values of each variable are linked visually) and residuals plots (map and scatterplot of residuals from a simple bivariate regression).

In “2D Scatterplot” you have to select your X and Y variables from the list of variables available, in the two boxes that list the variables in the left of the window. You can also select multiple Xs or multiple Ys by holding the CTRL button of your keyboard. In the Options tab (which only appears when one X and one Y variable are selected) you can easily fit linear and polynomial regressions directly in 2-D graphs. You can also check the value of each observation (may be an outlier, or a leverage) using the “Highlight Nearest Dot” option. You should explore all the features available under the “Options” tab.

As you should learn for every graph in SAM, the right-button of the mouse hides many options to customize and export your graph. Also, graphics in SAM can be double-clicked to open them in their own window (graphics opened in this way stay exactly as they are thereafter, whether or not you change data or settings). It is worth spending some time exploring the right-click menus because the graphs in SAM can be customized to publication quality. In the example shown, I used the “Edit Graph Labels” menu to change the size of the axis titles, as well as their respective labels.

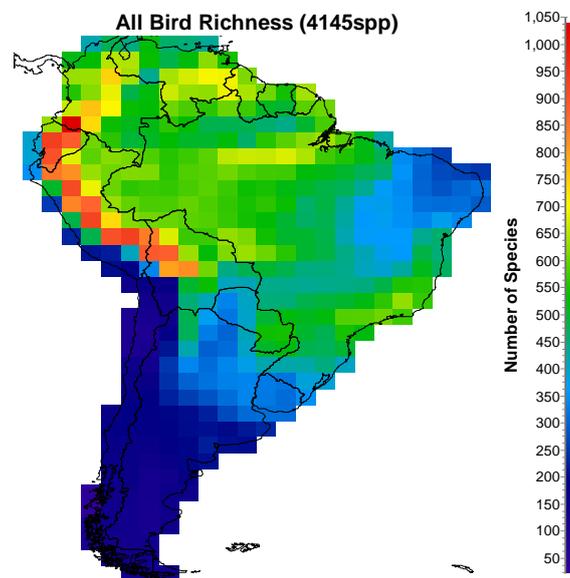


A map of the sort shown below is, of course, just a 2D scatterplot with a third variable represented by colors. In SAM v4.0 the maps are now highly customizable and allow publication-quality figures. Here is the “raw” map drawn by default by SAM for Mean Annual Temperature:



As you will find out, moving the mouse over the map shows the case number, coordinates and value of the variable in the cell hovered over. The most-used options in the right-click menu are: a) Edit color classes, where you can change the color palette used for your variable and the boundaries of each color (drag the colour bands on the histogram), b) Add map boundaries, c) Show or hide the color legend, d) Edit map labels, where you can change the text (font type, font size) associated with your map, e) Edit map legend, where you can change the text (font type, font size) of the legend of your map, f) Copy and Export figure, which allow you to use the figure in another program.

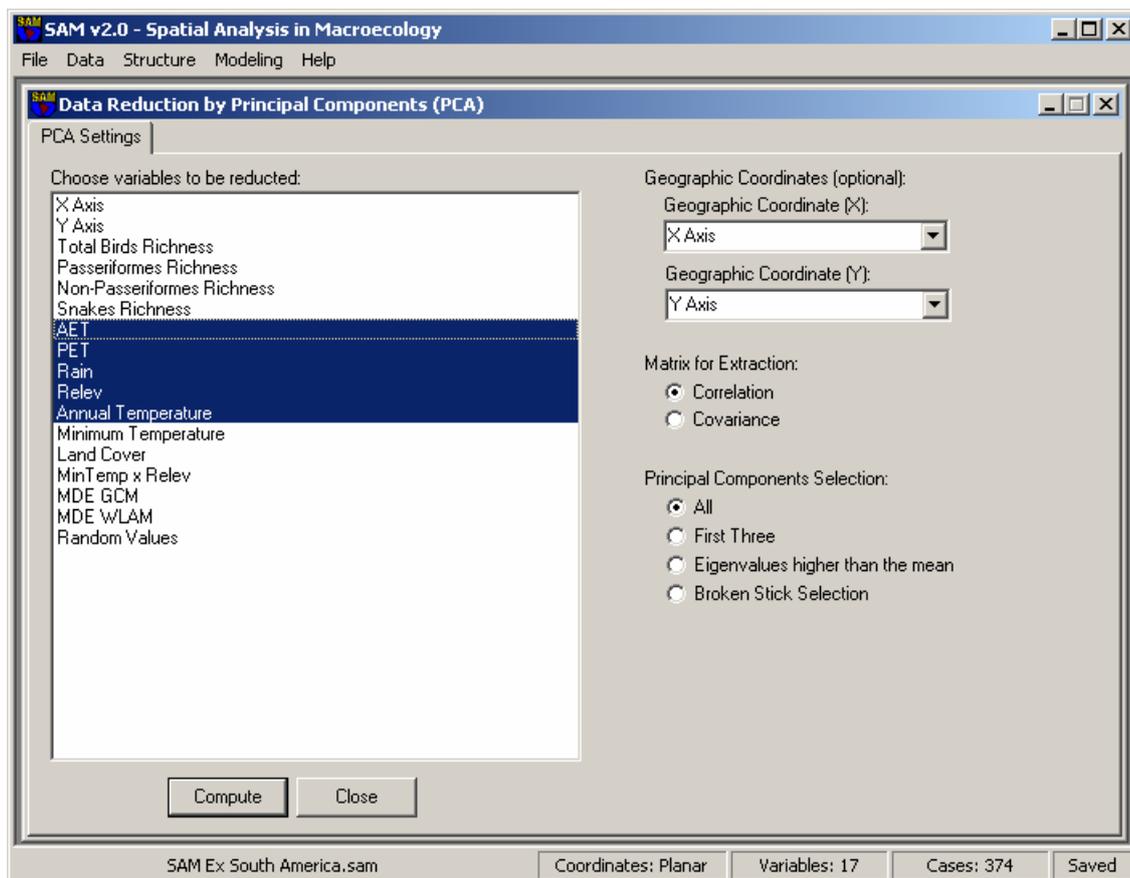
As an example of a customized map, the map below shows the pattern of species richness of all bird species, as calculated in the first steps of this tutorial:



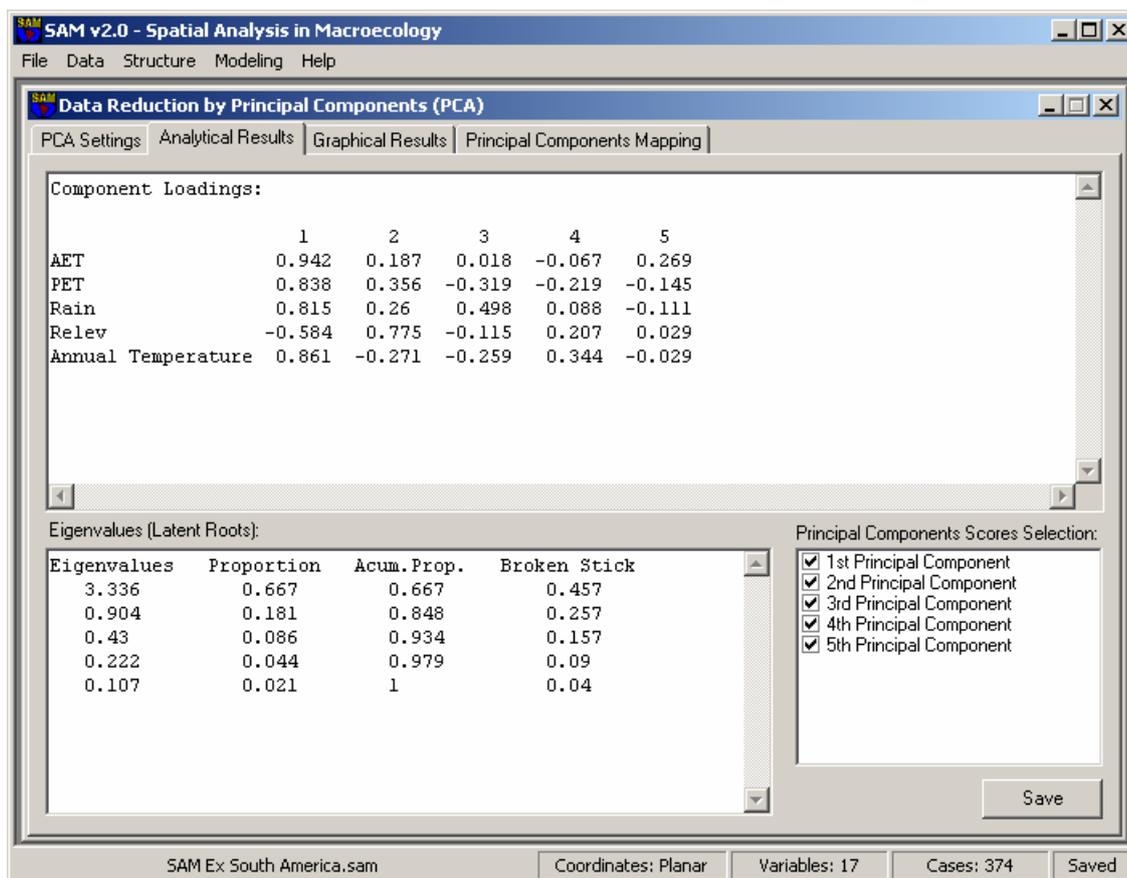
Principal Components Analysis

The main purpose of including PCA in SAM is to reduce co-linearity among explanatory variables (e.g. in a linear regression model). The module is quick and easy to use and is good for visualizing the output, especially mapping and examining the spatial structure of the resulting components. On the other hand, there are few options to select from (e.g. no option for axis rotation); if you want this extra functionality, or to perform factor analysis, you will need different software.

Data > Data Reduction (PCA). Select the variables to be reduced to a small number of orthogonal explanatory variables. (Note: the data used in the screen shots below are a different South American dataset: PET = potential evapotranspiration; AET = actual evapotranspiration; Relev = range in elevation within the cell.)



By clicking 'Compute', you get the first window of results:



Notice the “Save” button at the lower-right hand corner of the window. By clicking this you can save the PCA scores, and it will add them to your data for further analysis. This “Save” option is common to most of SAM modules, so you can save any output that is structured in the spatial units (i.e. the cases), such as spatial filters, model residuals, etc. **It is important to remember that when you click “Save” you only temporarily store the new variables** in the RAM memory of your computer; in order to permanently save the data in a file you should use the menu **File > Save**. You can also save all other data displayed in SAM modules by copying (CTRL+C) and pasting (CTRL+V) in the software of your preference (e.g. Notepad, MS Word or Excel).

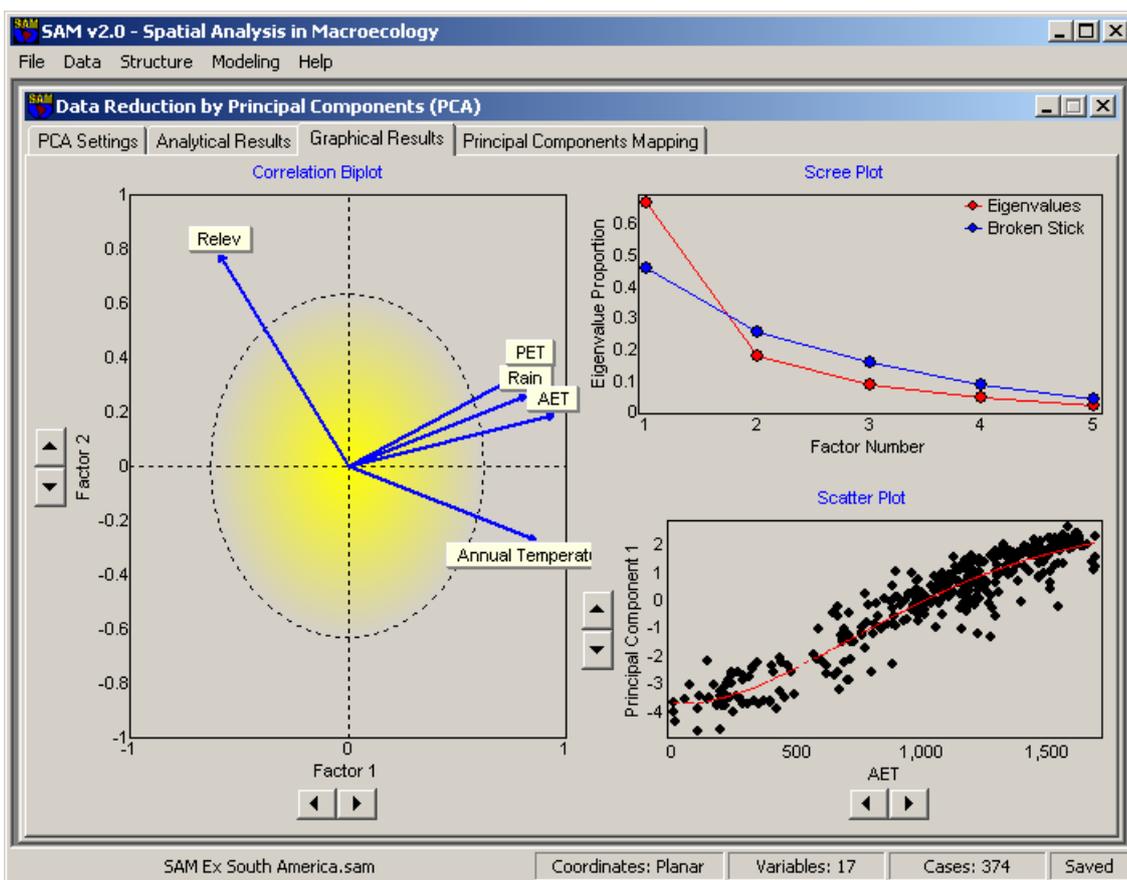
It is important to be able to find the following results:

- A)** the scree plot and broken-stick criterion, the loadings, the biplot and the variable vs component scatterplots, all of which aid the selection and interpretation of the components;
- B)** the maps and correlograms of the PCA scores.

In the “Analytical Results” tab, you will see the loadings (the contribution of each variable to each component), as well as the eigenvalues (left lower window). Notice that you can use the broken-stick criterion to select which axes are “important” or “significant”, by comparing the proportion of the variance accounted for (the second column) with the one expected by the

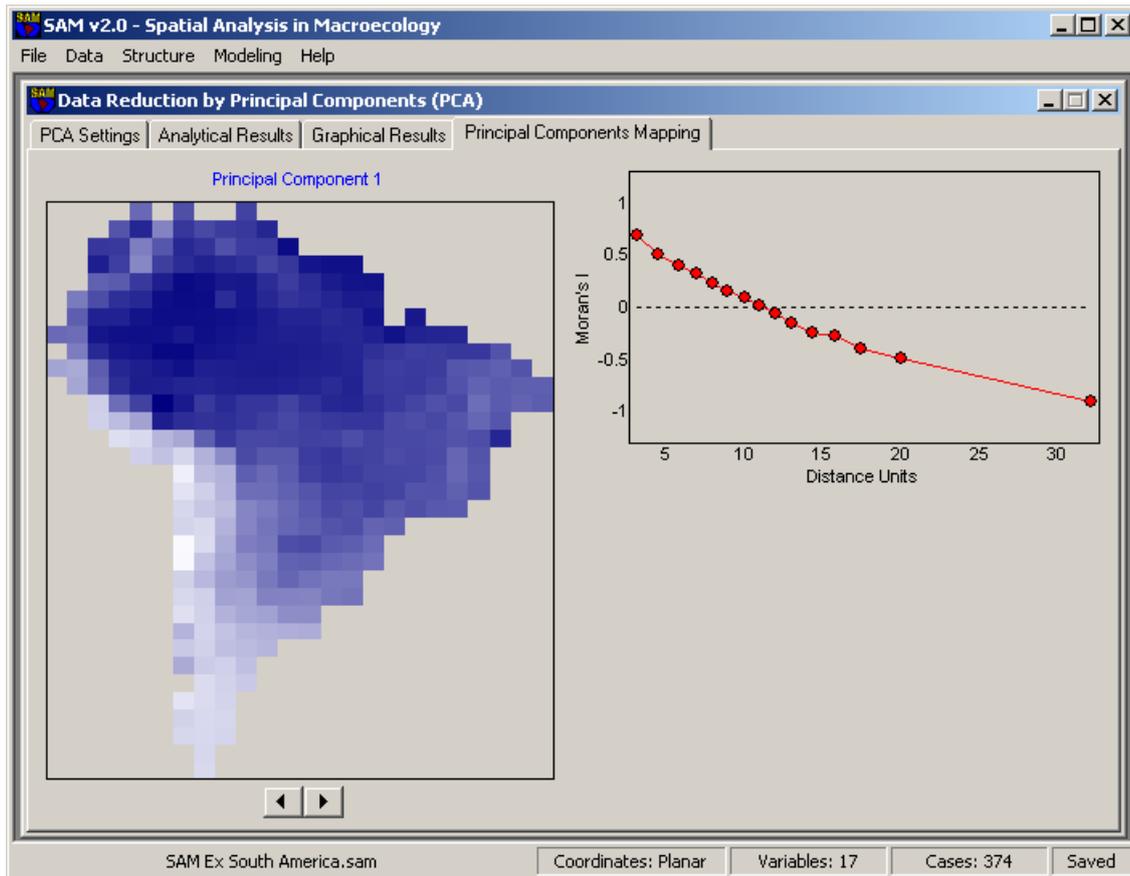
broken stick (last column). If the proportion of variation accounted for is less than that expected by broken stick, as in the example shown here, the axis would not be used, according to this criterion. Another commonly used criterion is the Kaiser criterion: selecting all axes with eigenvalues larger than 1 (absolute values, not relative).

Inspecting the loadings, you can see that all variables except Relev are highly positively correlated with the first component (for example, the loading of AET is 0.942). Relev has a moderate negative correlation with the first axis, but a strong positive one with the second. Notice that although only the first axis should be used according to the broken stick and Kaiser criteria, the PCA shows that Relev is distinctly different to the other environmental variables. Thus a third, very important criterion to select important axes is common sense! Axis 1 may be interpreted as “climate” and axis 2 as “topography”. In the analysis you have done, save the axes you have decided to retain.



You can visualize these results using the “Graphical Results” tab, in which the Correlation Biplot plots the component loadings. The yellow circle is known as “equilibrium circle”, an approximate rule of thumb in which any variable inside the circle would not be “significant” (none in the case shown). In the Scree Plot you can see the broken-stick selection: where the broken

stick is below the eigenvalue. Finally, the relationship between each original variable (e.g. "Rain") and the scores derived from each axis can be examined (click on the arrows to change the variable). Finally, you can also map of each new variable and see its spatial structure in a correlogram: click on the two arrows in the lower left-hand corner of the window to change the component. For more detailed spatial analysis, you can run other SAM modules on the saved components (see later exercises).



EXERCISES

From now on we have a series of exercises in statistical analysis and manipulation of data, with the main focus on spatial analysis – which SAM is particularly designed to facilitate. The techniques covered are all in the ‘Structure’ and ‘Modeling’ menus:

‘Structure’ menu: this includes several sub-modules for exploring spatial patterns in data, including spatial correlograms, trend surfaces, filtering, clustering and boundary delineation.

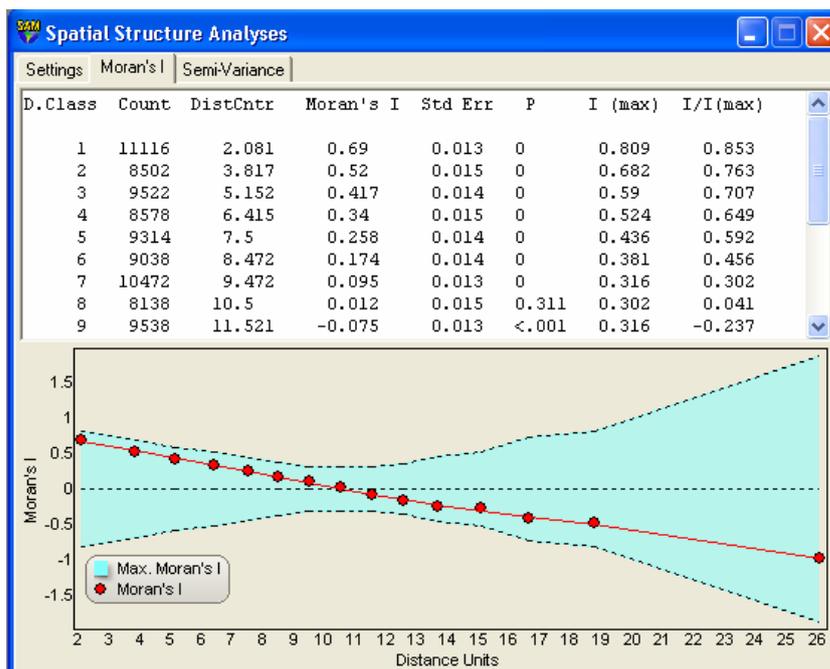
‘Modeling’ menu: this includes sub-modules for a wide range of spatial analyses, from simple spatial correlation to some of the latest and most complex spatial regression techniques.

The names of the exercises reflect the menu in which the relevant sub-modules are found (St for ‘Structure’ and Mo for ‘Modeling’).

Exercise St1: Spatial autocorrelation

Here we start analyzing the spatial structure of variables in the dataset.

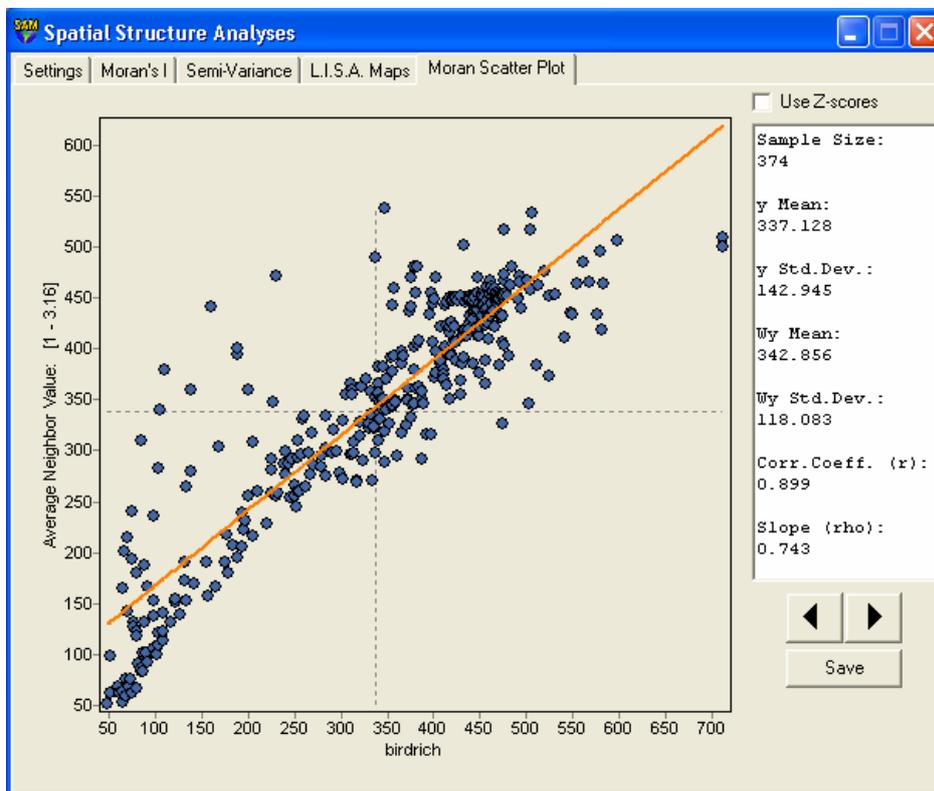
Press **Ctrl-i** or click  to bring up the **Spatial Structure Analyses dialog** (also Structure > Moran’s I and Correlogram). This is one of the core modules in SAM. Create spatial correlograms for different variables in the dataset. Try changing the number and type of distance classes. Interpret the output under the ‘Moran’s I’ tab: the columns in the correlogram output, which include the number of connections used in each distance class, the maximum distance, the Moran’s I and its error, the P-value and the maximum I.



This is the main screen of the output, showing Moran's I and the correlogram. You can see a pattern in bird species richness, with positive autocorrelation at short distances and negative autocorrelation at long distances. In the first distance class, the Moran's I was equal to 0.69 ± 0.013 ($P < 0.01$), and this value was calculated using the 11116 pairings of cells that are closest together (between 1 and 3.16 cells apart; 2.08 cells apart on average – here the grid system used is arbitrary, but if it were geodesic the units would be in km). The maximum possible value of Moran's I with this 'connectivity structure' is 0.809, and thus the relative value of Moran's I is 0.853.

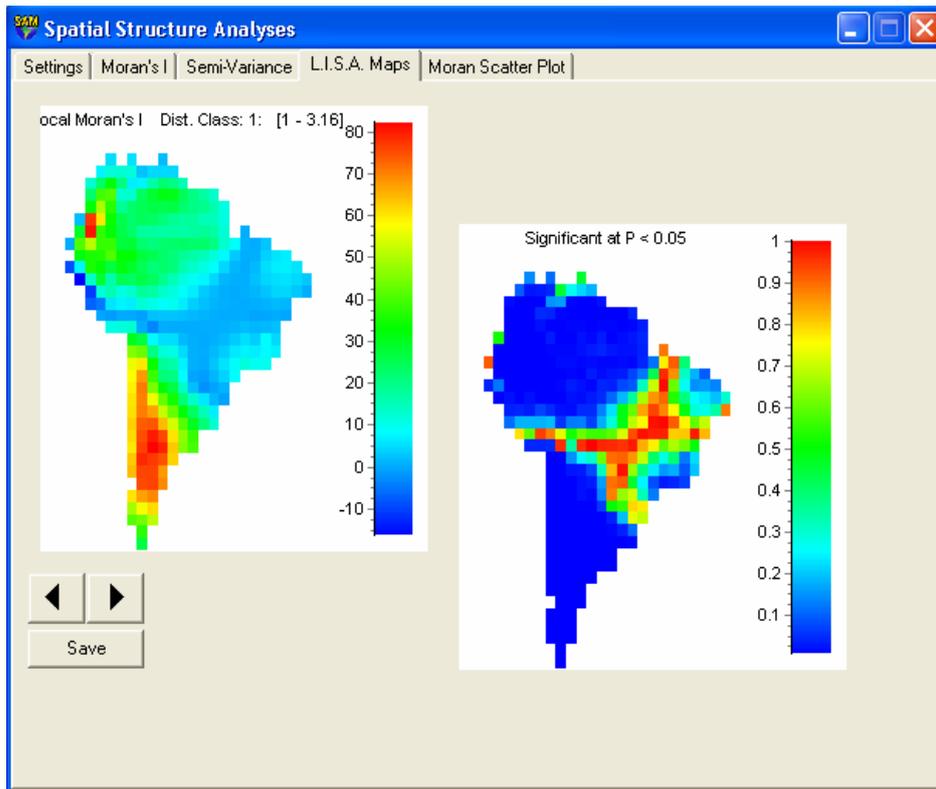
Generate the Moran Scatterplots. For each distance class (in turn) these show the relationship between the variable's values and those of the neighbouring cells.

You can see all the data, with dotted lines indicating the mean values. By default the screen shows the first distance class (as shown below); to see other distance classes click on the arrows. The distance classes are defined on the Y-axis label. Summary statistics are shown in the panel on the right.



Compute local autocorrelation (L.I.S.A.) and use Monte Carlo methods to test the statistical significance of the resulting Moran's I . This is good for analyzing regression residuals; however, it can be time consuming, so ensure you used no more than 200 randomizations ('permutations'). Interpret the maps of local autocorrelation that result, and their significance.

In 'L.I.S.A. Maps' you can see the maps of the local autocorrelation. In this case, there is clear overall autocorrelation structure, so local Moran's I can be interpreted as a 'leverage' statistic, showing the importance of each cell to the analysis of the overall pattern. NB you can save the local Moran's I values and their P-values.

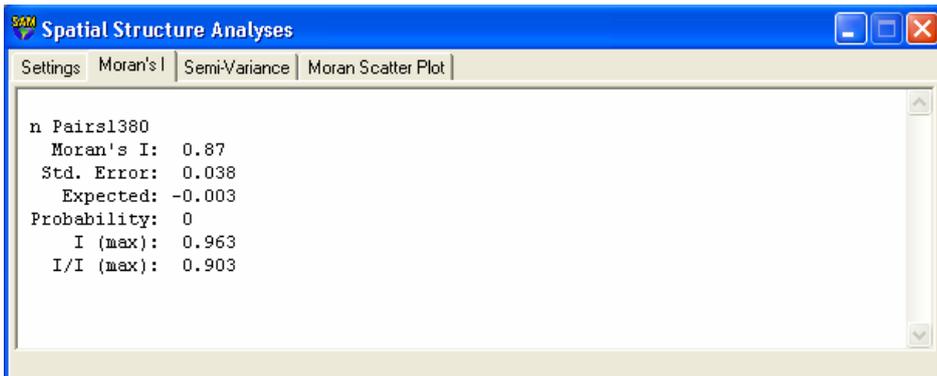
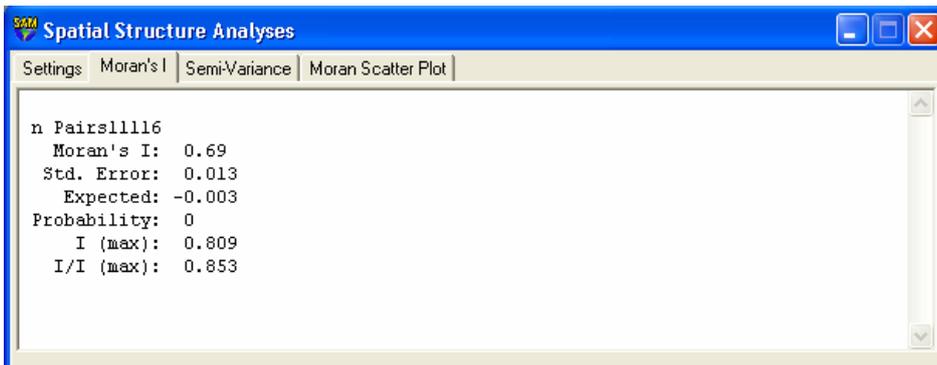


You can also explore the 'Semi-Variance' tab, trying different ways to fit alternative models and interpreting the diagnostic statistics. This is covered below when dealing with GLS in the 'Spatial Autoregression' submodule of SAM, so see below for more.

Exercise St2: Spatial autocorrelation using connectivity/distance matrices

Create two connectivity matrices using the 'Create/Edit Connectivity Matrix' procedure described above: Gabriel and a distance matrix that matches the smallest distance class in the previous exercise (in the example shown it is $0 < d < 3.17$). **Use these matrices in turn to obtain single Moran's I values** in the 'Spatial Structure Analyses' dialog (use the option "Use Available Connectivity Matrix", and then select the connectivity matrix you want).. This changes the way to define spatial relationships among cells, shifting from geographic distances to connectivity matrices.

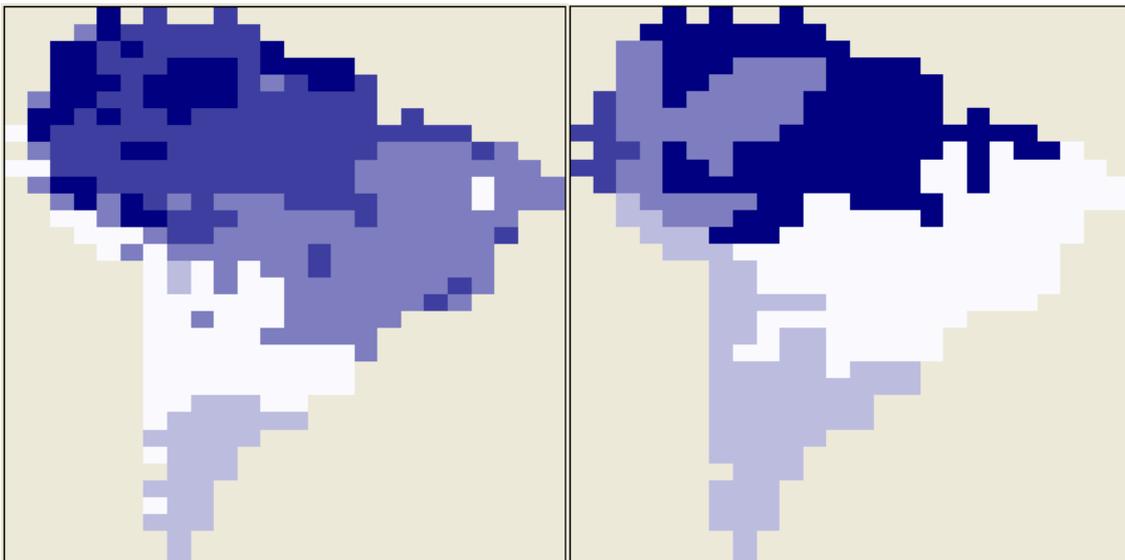
For the distance matrix you should find the same results as those in the previous exercise, for the first distance class (first screen grab below). To help understand correlograms better, you could generate the values for entire correlogram in this way. The results for the Gabriel matrix, which is calculated for only the very closest cells, show even stronger spatial autocorrelation (second screen grab).



Exercise St3: Spatial Clustering

This is K-means cluster analysis, with the option of spatially constraining the analysis.

Click  (or Structure > Cluster and Spatial Cluster) and do a **5-means cluster analysis of the five climate variables**, to make 'climatic regions' of South America. First do it without spatial constraint and then with spatial constraint, using one of your connectivity matrices. The example below uses the Gabriel criterion and Natural Breaks seeds.



Without spatial constraint

With spatial constraint

Cluster analysis by *K*-means is an iterative method to find the set of cells that minimize the TESS (Total Error Sum of Squares) within each group (here called “cluster”). Because there is no analytical solution to this method, the computational algorithm comprises an extensive search of possibilities, trying to optimize the solution. The algorithm starts by assigning one cell for each of the clusters, and then searching for the most similar cell to add to each cluster. When all cells are assigned to clusters, the next step is to iteratively move cells between clusters, each time seeing whether the overall TESS is reduced, thus maximize the similarity of the cells within each cluster. You can limit how long (number of iterations) the algorithm will try to improve the cluster analysis (minimizing the Sum of TESS among clusters) using the option “Maximum number of interactions” [complete with spelling error!].

Another option you have is how the cell seeding is done. The “Equal Intervals” option chooses cells equally spaced in the variable dimension. The “Natural Breaks” option ranks the cells according to a variable, and then chooses cells equally distant in this rank. You should be able to guess what the “Random” option does! The “Enter Seeds” option allows you to define the seed of each cluster.

So far, there is nothing “spatial” about clustering cells into groups. However, if you have a connectivity matrix you can spatially constrain which cell may join which group. The idea here is that all cells within a single cluster should be connected by the connectivity scheme you define. If your connectivity scheme connects only neighbors at close distances, then all the clusters in your analysis becomes continuous in space. However, once you impose a constraint to the optimization of the clusters, there should be more heterogeneity within clusters compared to unconstrained cluster analysis.

Be warned that for a large number of cells, and multiple variables at once, the procedure can be time consuming.

The analytical results tell you how many cells are within each group, which cell was the “seed” of each group, and the TESS of each group. Also, the “Sum TESS” among clusters measures how well the cluster algorithm could group similar cells, and “Mean TESS” can be used to show which group is more, or less, similar than the average.

Notice that using spatial constraint increases the Sum of TESS among clusters compared with an unconstrained analysis, because this is a less optimal solution to the problem. However, it is useful for defining geographic regions.

Before leaving, you can save a categorical variable that indicates to which cluster each cell belongs.

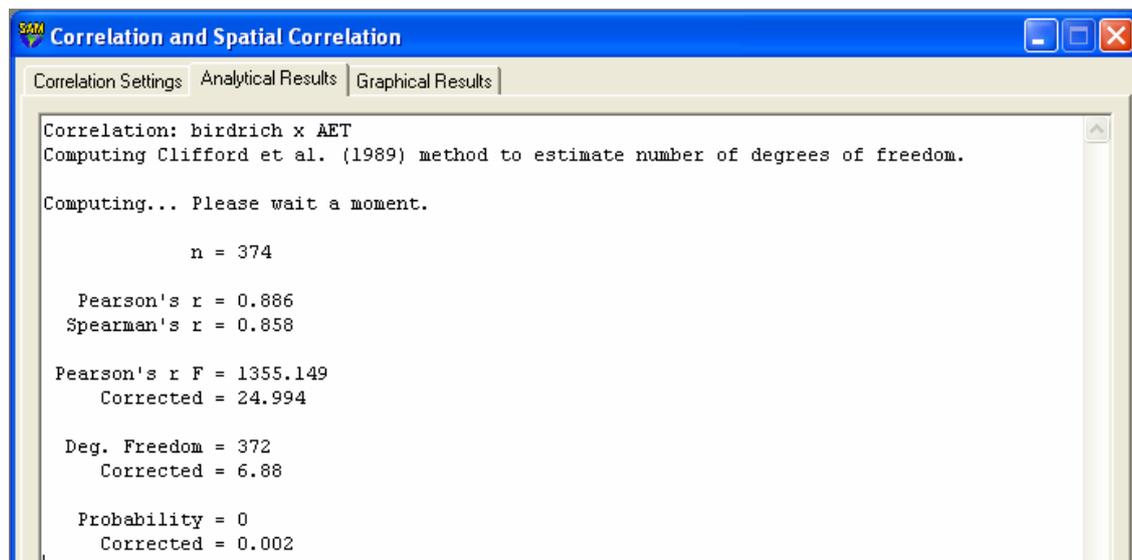
Exercise Mo1: Spatial Correlation

Here we see how to do correlations that are 'corrected' for spatial autocorrelation (taking into account the magnitude of spatial autocorrelation in two variables).

Press **Ctrl-N** or click  to bring up the **Correlation and Spatial Correlation dialog**. First, choose no geographic co-ordinates and calculate a matrix of Pearson correlations between all the climate variables and bird species richness.

Choose X and Y as the geographic co-ordinates and **calculate the correlation between two of the variables, using the procedures to correct degrees of freedom, via two different estimators (Dutilleul and Clifford)**. NB with spatial coordinates (the default option), you are only allowed to choose two variables.

Choose one of the methods (Clifford is faster) and see how varying truncation distances affects the geographically effective degrees of freedom and F and P.



With the example shown, if you choose bird richness and AET (using the CTRL key) the spatial correlograms of the two variables will be displayed. After computing, you will see that the degrees of freedom drop from 372 ($n - 2$) to only 6.9, using Clifford's method! Even so, the correlation of 0.885 is still significant ($P < 0.001$). If you use a truncation distance of 11, only the positive parts of the correlogram are used, and the DF become 13.8 – this is a bit more liberal, and assumes that Type I error problems are caused only by short distance spatial autocorrelation; long distance autocorrelation, at least in this case, is better interpreted as spatial dependence.

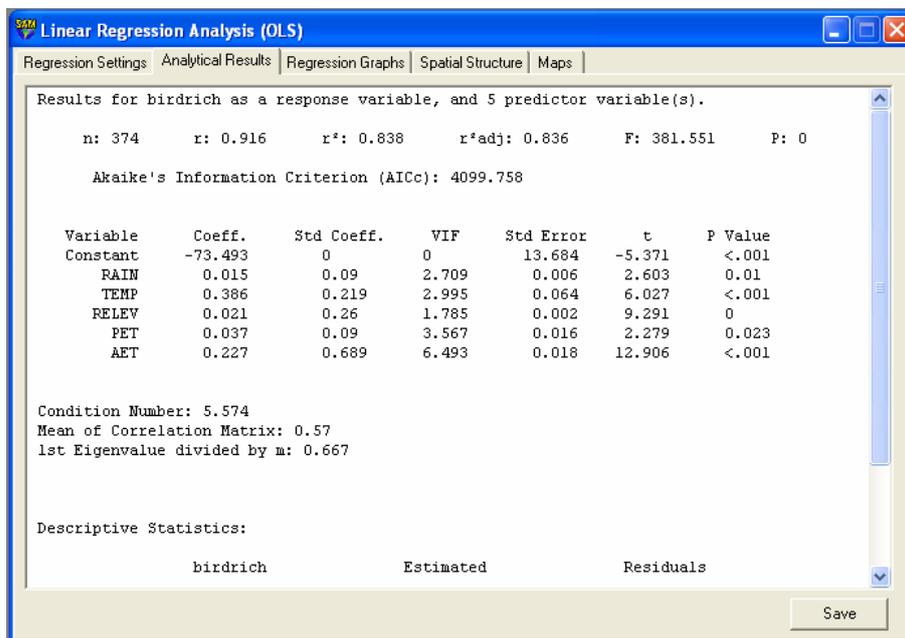
Also examine the graphical results. The map shows the departure from the best-fit line (left-hand graph) spatially.

Exercise Mo2: OLS linear regression

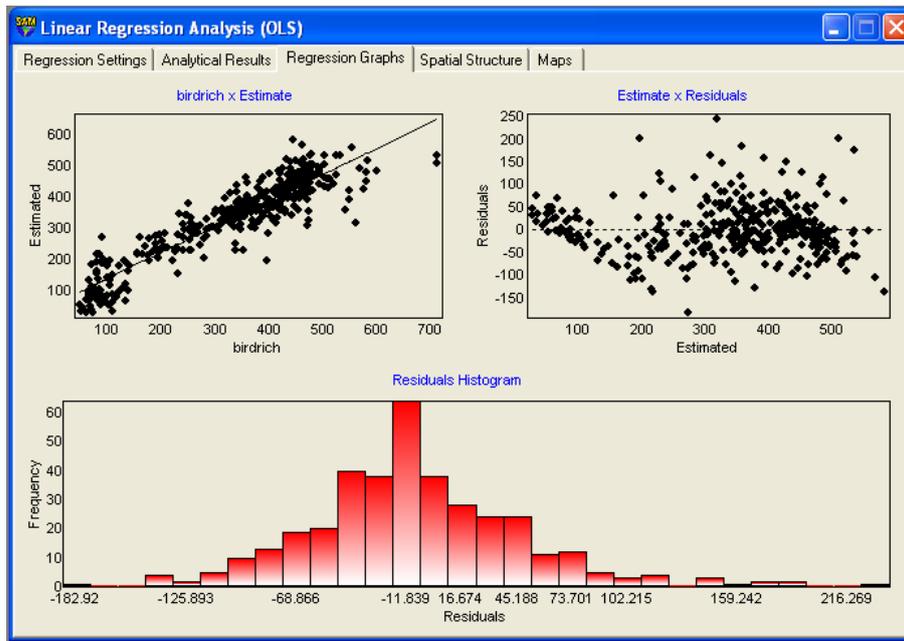
This is just the normal, non-spatial regression.

Press **Ctrl-R** or click  to bring up the **Linear Regression Analysis dialog**. The analytical output in SAM is the standard regression output; pay attention to coefficients of determination, AIC values and partial slopes, with their associated statistics (including the variance inflation factor, which tests for multicollinearity). The final three tabs that appear allow you to perform regression diagnostics, including analysis of the spatial structure in the residuals. Check how including different predictors changes the level of residual spatial autocorrelation. Note that this is only for immediate diagnosis; a better evaluation of patterns in the residuals (including significance tests of Moran's *I*) should be done on saved residuals, via the 'Spatial Autocorrelation Analysis' sub-module, as described in Exercise St1.

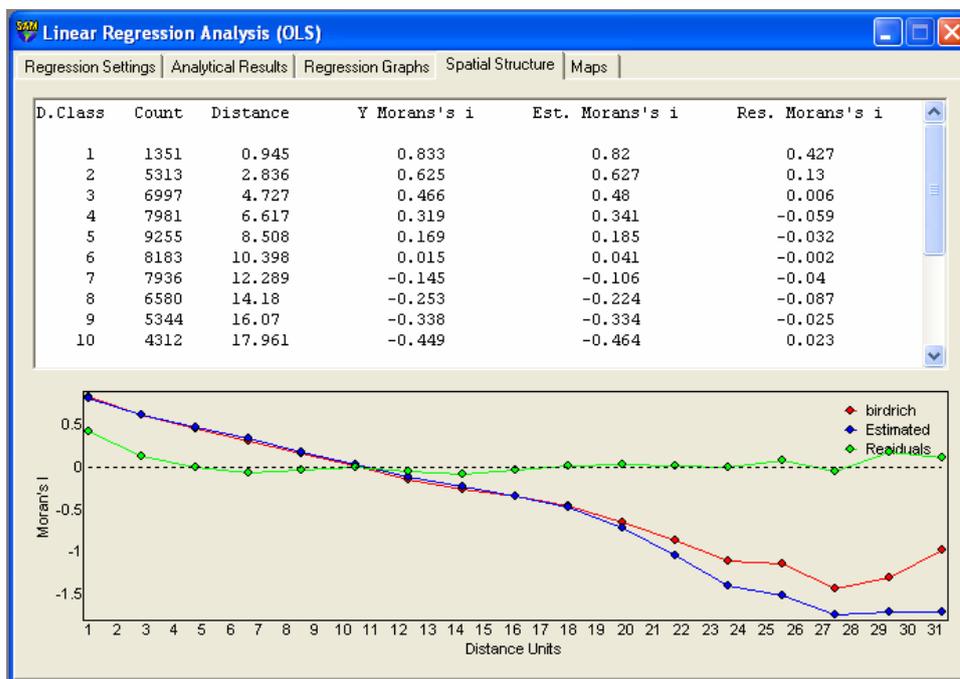
This is a very important module, and is explored further in the next few exercises. If you run a standard OLS model of richness against the five environmental predictors in the example shown, you get the following analytical results. The five environmental variables account for 83.8% of the variation in richness (83.6% using the adjusted R^2). The most important explanatory variable is AET, according to the standardized regression coefficient (= 0.69). All explanatory variables are significant at $P < 0.01$, except PET; however, these P-values may well be biased by the presence of residual spatial autocorrelation.

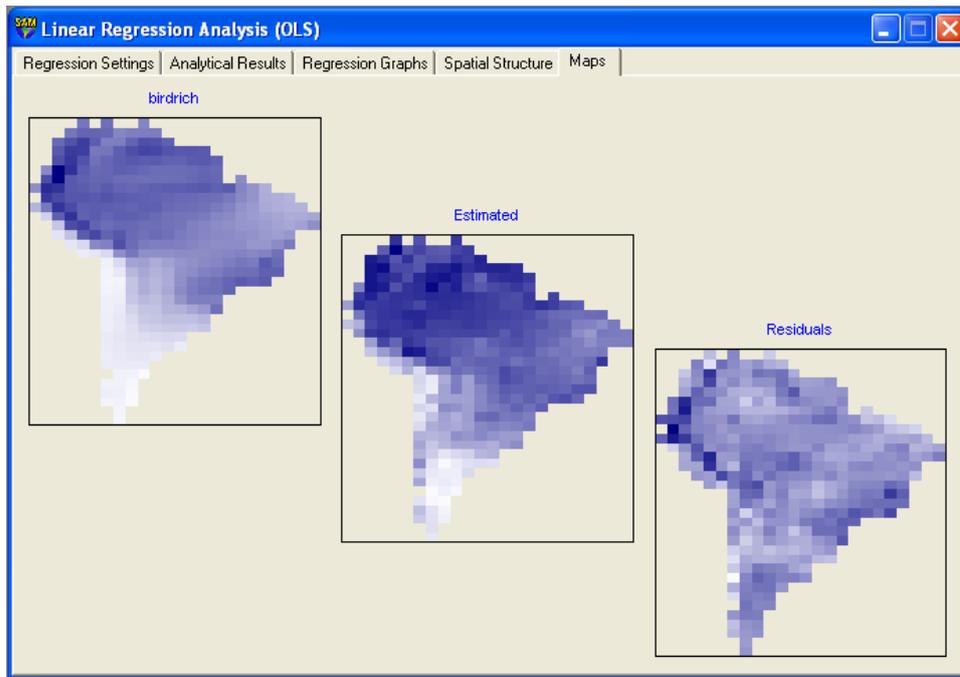


The next screen (regression graphs), contains standard residual diagnosis graphs, including the histogram of the residuals, which seem to be non-normal, leptokurtic – you can check the skewness and kurtosis values in the previous screen, in the column “Residuals”, here equal to 0.613 and 1.874. You also have a plot of the observed versus the predicted values (you can check for linearity here) and a plot of the residuals versus predicted values (no pattern is expected here).



Perhaps the most interesting screens (in terms of what SAM does compared with most software) are the next two, which allow you to explore spatial patterns in the model, especially in model residuals (the green line of the correlogram). There is a relatively high Moran's I of 0.25 in the first distance class, indicating that Type I errors are biased. More important, there is a lack of explanation of richness at these short distances. Look at the maps of the residuals (last screen) and think about what is lacking in the model.

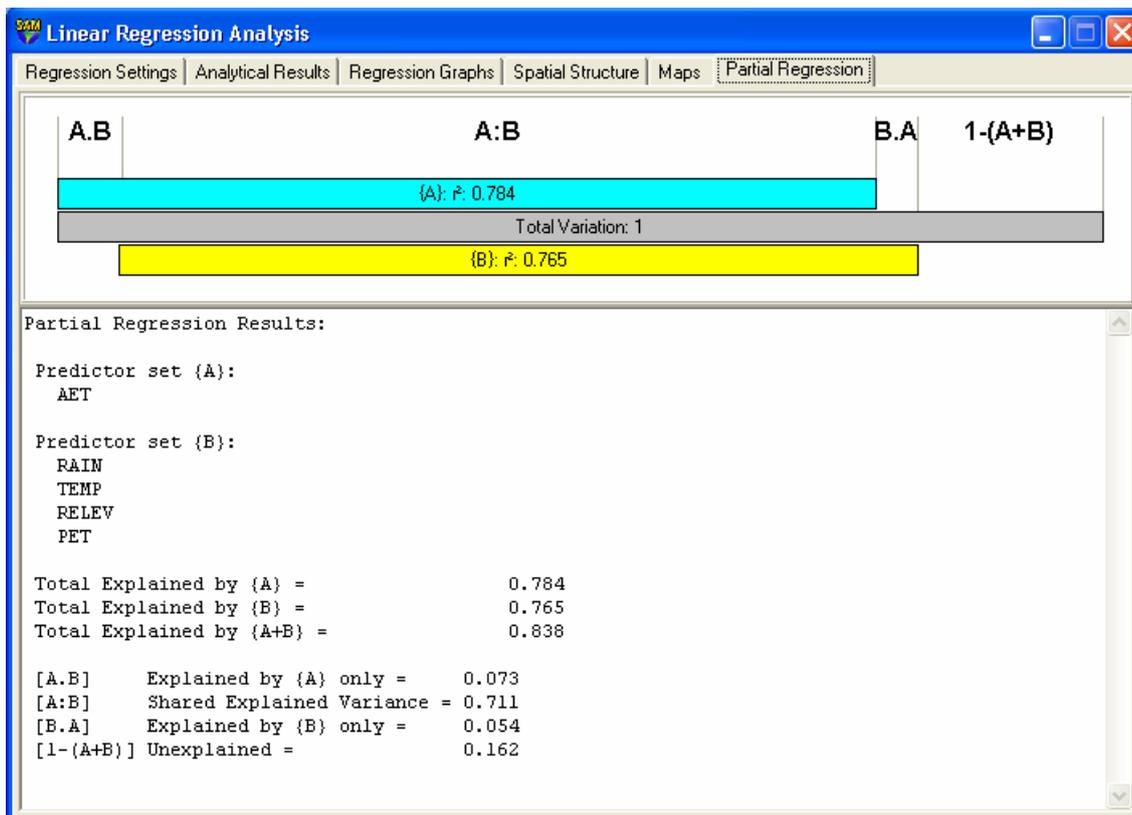




Exercise Mo3: Partial regression

Still using OLS regression, we now define ‘predictor sets’ to examine overlap in ‘explanation’.

Go back to the ‘Regression Settings’ tab. Enter some of the X-variables in one ‘predictor’ set and the others in another (use the button with a + in it, to the right of the box for ‘Sets of Predictors for Partial Regression’ to create a second set of explanatory variables). You can separate the explanatory variables in any way you wish – the example shown below demonstrates the issue of multicollinearity when AET is defined as one predictor set and four other climatic variables as the other. More commonly you may wish to separate spatial variables from environmental ones, which we return to in later exercises.



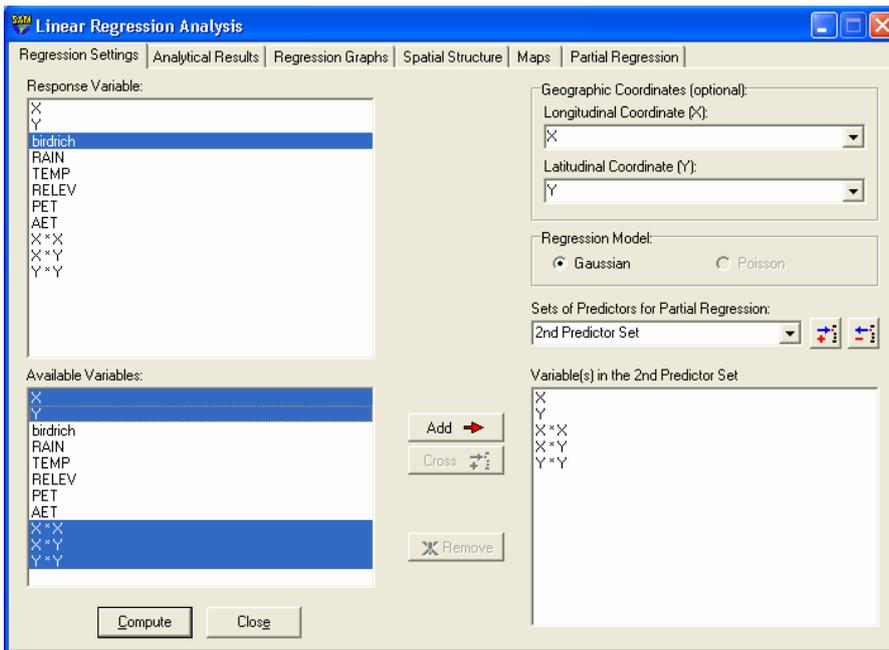
You can see the R^2 of each sub-model (AET only, other variables only and both), as well as the partition into unique and shared components. Only about 7.3% of the variation in bird species richness is explained solely by AET, and only 5.4% solely by the other four variables together. By far the most of the explanation (71.1%) is shared, reflecting the strong multicollinearity in the environmental variables. Note that this multicollinearity is also measured in the 'Analytical Results' screen by the variance inflation factors (one for each X-variable) and the condition number (like an overall VIF for the model).

Exercise Mo4: using a Trend Surface

This time we do a spatial partial regression, using a trend surface as the spatial 'predictor set'.

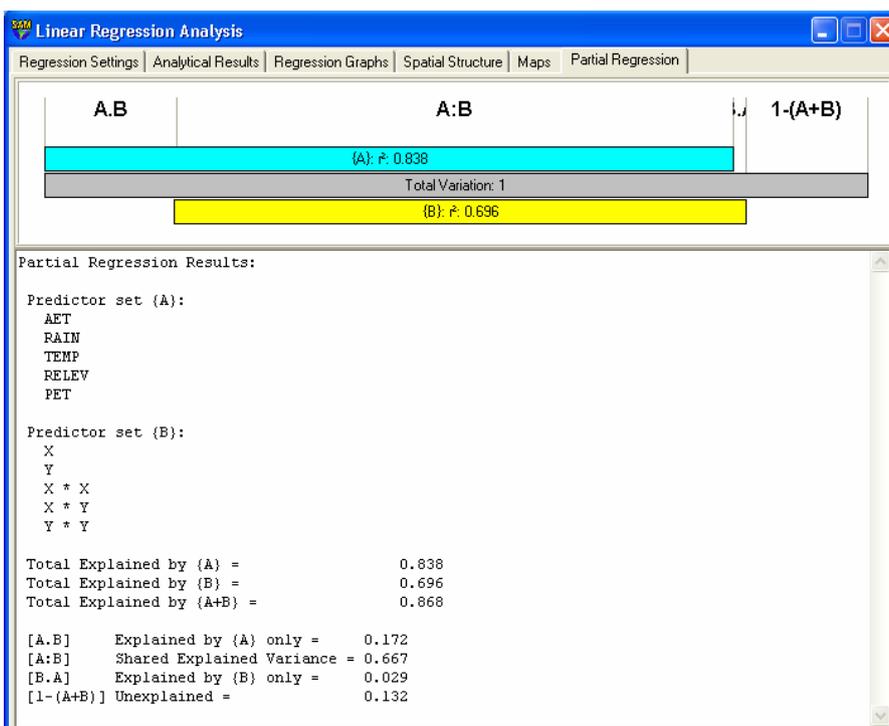
In Data > Data Handling > Polynomial Expansions, calculate the variables for a second-order trend surface. These are simply your spatial variables (X and Y or longitude and latitude – select these in the left panel) and their polynomial expansion (order: 2). Generate the polynomial terms and then Save them by moving them from the list of new temporary variables into the main dataset.

Then enter the five spatial variables as one 'predictor' set and the environmental variables in a second. Notice how adding broad-scale spatial variables reduces the residual autocorrelation only a little.



In the example shown, using the a 2nd-order trend surface ($X + Y + X^2 + XY + Y^2$) increases the model R^2 to 0.868. Again the partial regression is dominated by the overlap, this time between 'space' and 'environment'. About 17.2% of the variation in bird species richness is explained solely by the environmental variables, and only 2.9% by 'pure' spatial variation. However, most of the explanation (66.7%) is shared, reflecting spatially structured environmental variation.

After including the spatial variables, AET is still the most 'important' environmental variable (see the standardized coefficients in the 'Analytical Results' tab). Note also the standardized coefficient for latitude. The residual spatial autocorrelation in the first distance class only dropped from 0.427 to 0.407, because only broad-scale trends were used. The maps of the residuals are similar to the original OLS that included only environmental predictors.



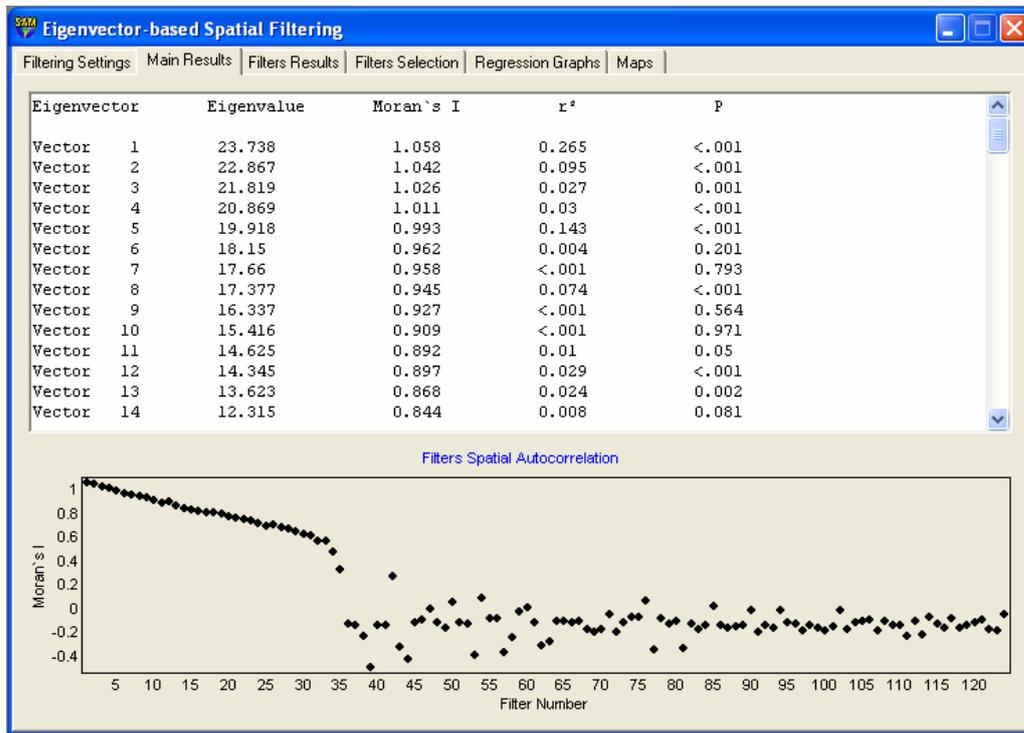
You can use AIC to try different model combinations, selecting the one with smaller AIC (differences of 3 or larger are often seen as enough to indicate better model fit). Warning: since residual variance is downward estimated (because of spatial autocorrelation), this criterion will be 'biased' and there will be a tendency to select over-complex models (with many variables).

Exercise Mo5: Eigenvector-based spatial filtering

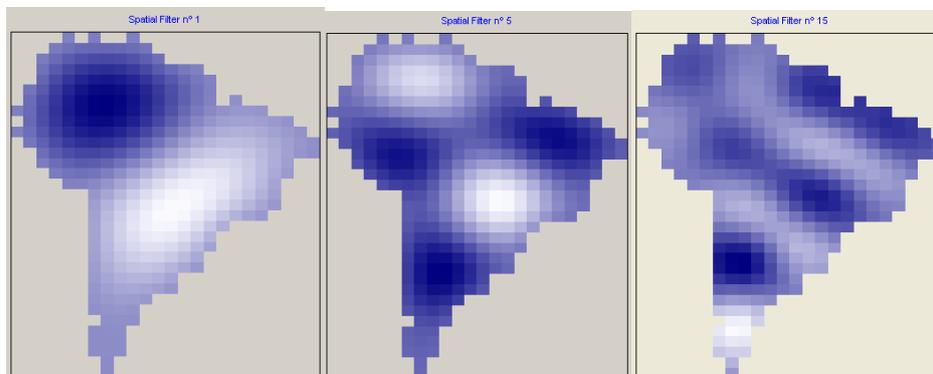
Also known as "Spatial EigenVector Mapping" (SEVM), this is a technique that has been quite recently developed and shows some promise. The purpose of this exercise is to use SEVM to remove spatial trends in a (response) variable. You should notice that spatial pattern changes among filters, but in general becomes less and less structured as the eigenvalue associated with the filter becomes smaller.

Press Ctrl-F or click  to bring up the Eigenvector-based spatial filtering dialog. Explore the routine for eigenvector-based spatial filtering, interpreting the outputs. Note (a) the need to establish a truncation distance and (b) that choosing the response variable is only necessary to help in diagnosis (it is not intrinsically part of the filtering procedure). The challenge here is how to select filters for further analyses – an issue not resolved in the literature, but minimizing residual short-distance spatial autocorrelation is usually a good idea. Observe the spatial patterns of each filter (maps) and recognize that spatial structure starts to become more 'local' when using filters with smaller eigenvalues. Notice how selecting filters will reduce the level of residual autocorrelation. Save the filters you selected, for further analyses.

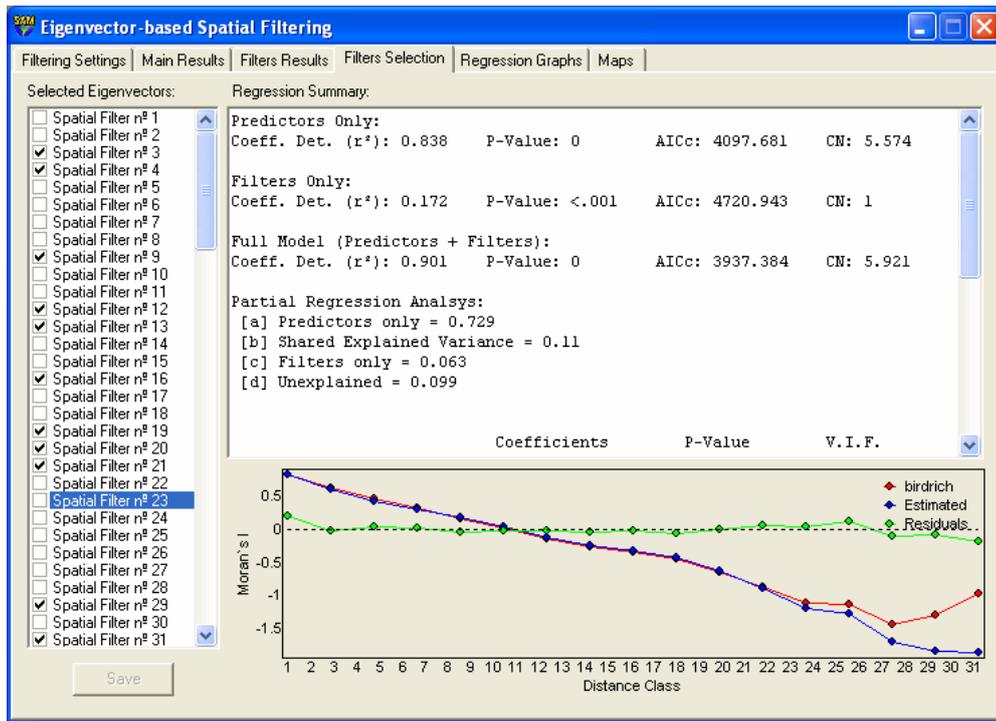
In the example shown, bird richness is the response variable (optional – just to help diagnosis), the five environmental variables are 'predictors' in a partial regression (again optional) and a truncation distance of 5 is used. You will see the eigenvalues and the Moran's I, in the first distance class, for each filter. The first filter has a Moran's I a bit higher than 1 (1.06), and a squared correlation of 0.265 with richness. Notice that the figure is NOT a correlogram, but it shows the short-distance Moran's I for all the filters.



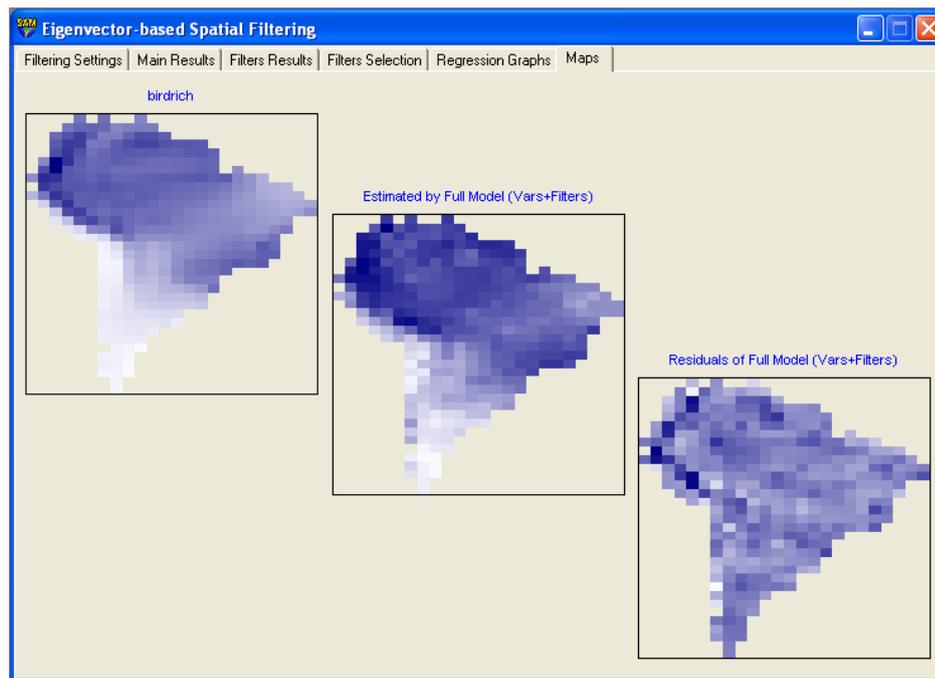
In the second screen, you can see the maps and other graphics for each filter, including the correlogram and the relationship with the response variable you chose. It is interesting to see spatial structure changing when obtaining filters with lower eigenvalues, as shown below for filters 1, 5 and 15. This structure is also dependent on the truncation distance (which explains why there are two patches, not one, in the filter 1).



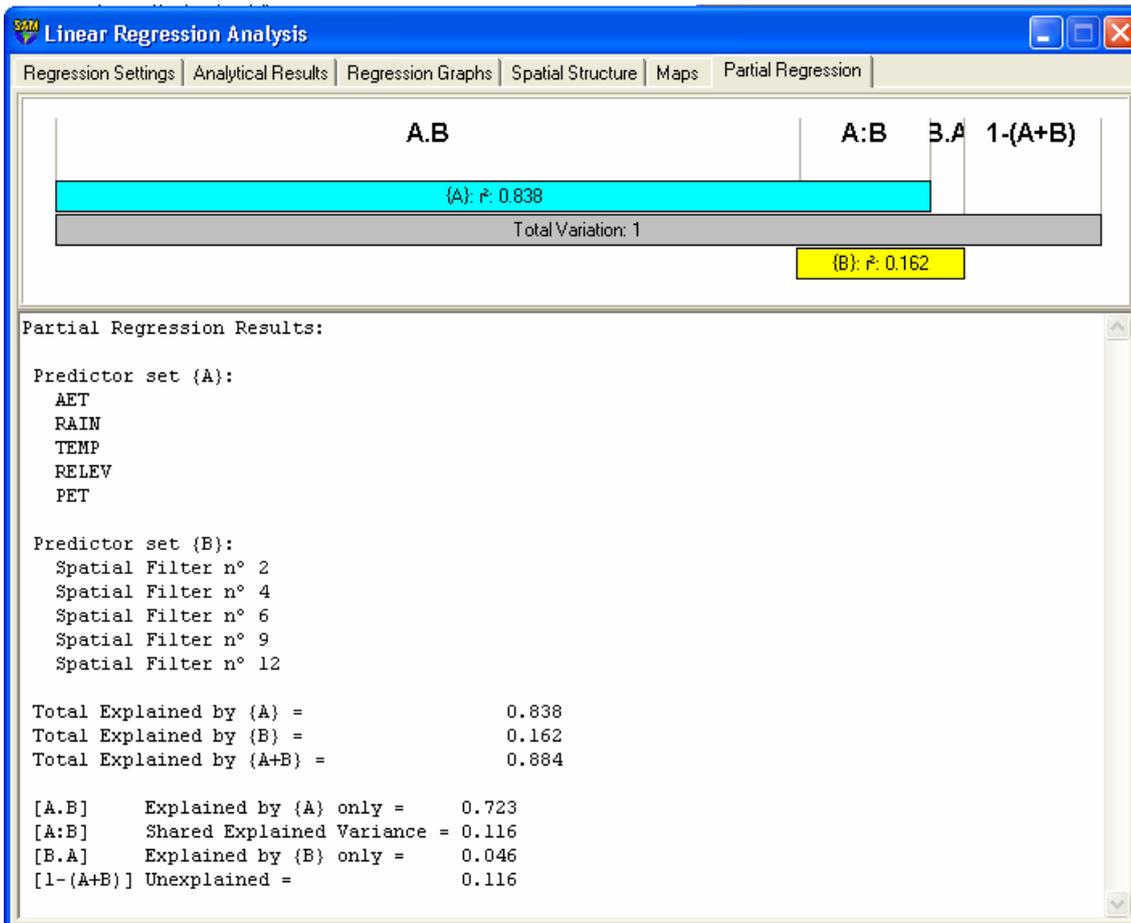
The next screen is to help selecting filters. The filters marked respect the criterion you used in the opening screen, but you can add or delete other filters. Each time you select a filter, it will update the parameters of the multiple regression, the correlograms and the maps (in the next screen).



For example, using the first 5 filters will give you an squared-correlation with bird species richness of 0.56, but there is still a significant (almost 0.5) Moran's I residual autocorrelation in the short distances. Increasing to the first 10 filters, the R^2 goes to 0.64, but quite a lot of residual spatial autocorrelation still persists. A reasonable balance is achieved by choosing filters 3, 4, 9, 12, 13, 16, 19, 20, 21, 29 & 31 – this reduces residual spatial autocorrelation to a low level, while not overlapping much with the environmental variables in partial regression analysis. But looking at the maps in the next screen (and below) you can still see some local structures in the region of Andes. Choosing filters 2, 4, 6, 9 and 12 is simpler and quite satisfactory but leaves a bit more small-scale spatial autocorrelation in the residuals; again the map of residuals shows patterning in the tropical Andes. Filter selection is a difficult issue!



Save your filters, and also the “Linear Combination Filters” for further analysis and Press **Ctrl-R** or click  to bring up the Linear Regression Analysis dialog again. Use the filters you selected to evaluate, using a partial regression within the OLS routine, how adding the filters as independent variables affects the outputs, especially the partial regression coefficients, the coefficient of determination and, more importantly, the residual autocorrelation. This will add a little to what you have just done. Repeat the process using the linear combination of filters (which is a single variable representing the selected filters) instead of the filters.



Exercise Mo6: Multi-model inference

This is an increasingly popular and rigorous approach to model selection.

Press **Ctrl-G** or click  to bring up the Multi-model inference dialog. Using the five environmental variables, choose a model for bird species richness.

Now add the five environmental variables as variables to be selected, and the linear combination of filters as a variable (or all the filters you selected as variables) to be present in all

models. Note that SAM tells you how many models (i.e. possible combinations of variables) will be evaluated; if this number gets very high, computational time will be considerable.

Model Selection and Multi-Model Inference

Model Settings | Model Selection Results | Model Averaging Results | Model Averaging Graphical Results

Total of 31 OLS models run, in <.01 minutes

Response variable is birdrich.
Explanatory variables are:
Variable #1: RAIN
Variable #2: TEMP
Variable #3: RELEV
Variable #4: PET
Variable #5: AET
Variable #6: Linear Combination Fi...

Results for the OLS Model Selection procedure, sorted by Akaike Information Criterion (AICc).

Model	Variables	nVars	r ²	Cond.Num.	AICc	Delta AICc	L(gi x)	AICc wi
Mod #5	1, 2, 3, 4, 5, 6	6	0.884	6.013	3976.666	0	1	0.664
Mod #20	2, 3, 4, 5, 6	5	0.883	5.039	3978.03	1.364	0.506	0.336
Mod #8	1, 2, 4, 5, 6	5	0.869	5.561	4021.569	44.903	<.001	<.001
Mod #23	2, 4, 5, 6	4	0.868	4.454	4022.001	45.335	<.001	<.001
Mod #21	2, 3, 5, 6	4	0.866	3.253	4028.51	51.844	<.001	<.001
Mod #6	1, 2, 3, 5, 6	5	0.866	4.292	4028.68	52.014	<.001	<.001
Mod #27	3, 4, 5, 6	4	0.862	4.368	4039.561	62.894	<.001	<.001
Mod #12	1, 3, 4, 5, 6	5	0.862	5.35	4039.904	63.238	<.001	<.001
Mod #30	4, 5, 6	3	0.858	3.577	4047.344	70.678	<.001	<.001
Mod #15	1, 4, 5, 6	4	0.859	4.648	4047.725	71.059	<.001	<.001
Mod #4	1, 2, 3, 4, 6	5	0.842	3.621	4090.189	113.522	<.001	<.001
Mod #28	3, 5, 6	3	0.836	1.706	4101.312	124.646	<.001	<.001
Mod #13	1, 3, 5, 6	4	0.836	3.276	4102.809	126.143	<.001	<.001
Mod #7	1, 2, 4, 6	4	0.836	2.734	4103.966	127.3	<.001	<.001

Model Selection and Multi-Model Inference

Model Settings | Model Selection Results | Model Averaging Results | Model Averaging Graphical Results

Parameter estimates averaged across 31 OLS models, using Akaike Weights (AICc wi)

n: 374 r: 0.94 r²: 0.884 r² Adj: 0.883 AICc: 3975.256

Variable	Importance	Coeff.	Std Coeff.	Std Error	t	95% Lower	95% Upper
Constant	-	-119.812	0	12.219	-9.806	-143.761	-95.675
RAIN	0.664	0.009	0.054	0.003	2.778	0.003	0.018
TEMP	1	0.455	0.258	0.055	8.324	0.348	0.564
RELEV	1	0.014	0.173	0.002	6.986	0.01	0.018
PET	1	0.114	0.279	0.015	7.515	0.084	0.144
AET	1	0.183	0.557	0.016	11.163	0.151	0.208
Linear Combination Fi...	1	1.005	0.259	0.083	12.102	0.842	1.162

Parameter estimates for the best OLS model, according to the Akaike Information Criterion (AICc).

Variable	Coeff.	Std Coeff.	Std Error	t	95% Lower	95% Upper
Constant	-119.593	0	12.203	-9.8	-143.511	-95.675
RAIN	0.009	0.054	0.005	1.845	<.001	0.018
TEMP	0.457	0.259	0.055	8.369	0.35	0.564
RELEV	0.014	0.174	0.002	7.01	0.01	0.018
PET	0.115	0.28	0.015	7.558	0.085	0.144
AET	0.178	0.541	0.015	11.53	0.148	0.208
Linear Combination Fi...	1	0.258	0.083	12.079	0.838	1.162

This suggests that all the variables are important. However, note that we have not completely eliminated the small-scale spatial autocorrelation, so the routine has probably over-fitted. The selection and model averaging suggest that AET is the most important variable by far, in accounting for bird species richness. The condition number (a kind of aggregate VIF score) suggests issues of multicollinearity with many of the models, but notably not for models 31 (AET + filters) and 28 (AET + topography + filters). These simple models would be worth further

investigation. Model 28 is very similar to a model resulting from a PCA of the five environmental variables: PCA1 + PCA2 + filters – try this if you like.

Exercise Mo7: Spatial autoregression – Generalized Least Squares

This exercise and the next one look at various forms of autoregression, a commonly used form of spatial regression. This involves three groups of techniques that can all be found under Modeling—Spatial Autoregression. This is a very complex set of routines, although all outputs are similar to the OLS previously described.

Press Ctrl-U or click  to bring up the Lagged Models dialog. The ‘lagged-models’ are based on the residual analyses of ‘pure autoregressive’ models. Notice the effect of alpha and how autoregression, using the response variable, eliminates the residual spatial autocorrelation.

In the ‘Pure autoregressive model’, SAM allows you only to choose a response variable, because this first model deals only with a single variable. In the first screen of results are the r^2 and the AIC; the new parameter estimated, the “spatial autoregressive coefficient”, is here equal to 0.885 ± 0.043 for richness. So, according to r^2 , about 79% of the variation in bird species richness can be accounted for by the spatial weighting structure (1/geographic distances) among cells. The other screens are similar to OLS, and notice that the Moran’s I is 0.237, so this simple autoregressive model still leaves short-distance spatial structure, and the map in the last window will show why (again)! If you go back to the opening (“Settings”) screen and mark the “Lagged Response” or “Lagged predictor” models you can repeat the procedure, but incorporating environmental ‘predictors’ too. But let’s wait for the autoregressive models in the last routine (SAR and CAR), to try this.

Click  to bring up the Generalized Least Squares dialog. This runs an iterative form of Generalized Least Squares (GLS), also called ‘kriging regression’ or ‘generalized kriging’, in which the residual autocorrelation in the OLS is modelled by semi-variograms (see Exercise St1). These patterns are incorporated into the GLS procedure to obtain all coefficients. This routine is useful for understanding the logic underlying spatial regression and is a good one to focus on. The last group of routines includes the well-known conditional and simultaneous autoregressive models (CAR and SAR). In general, the outputs are similar, and for learning purposes, any one can be used. Pay attention to the following:

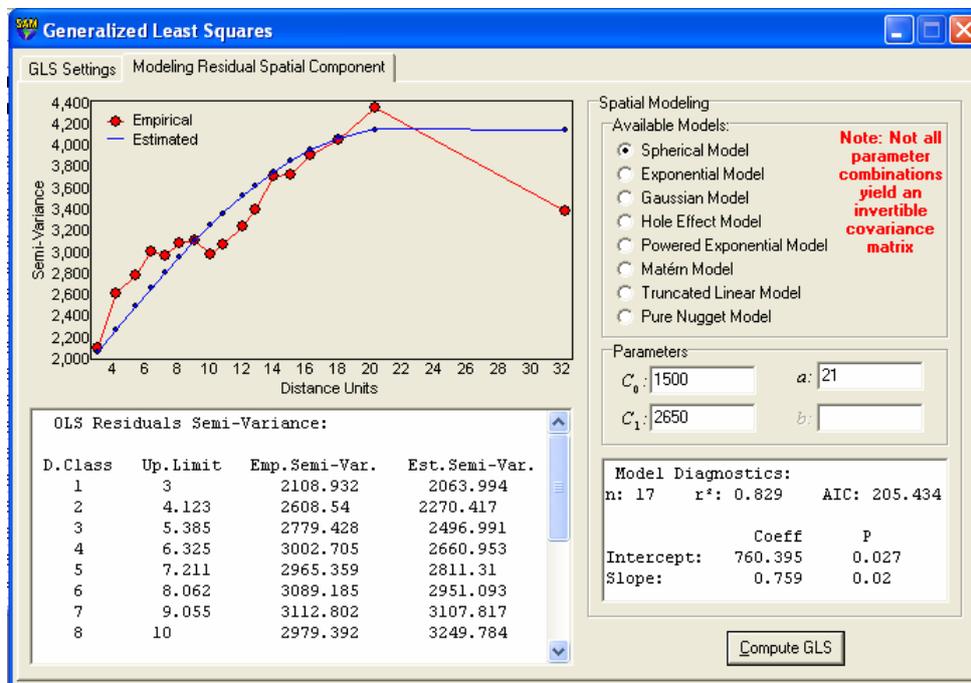
- The existence of an autoregressive coefficient, which can be calculated by the procedure or can be entered directly (in SAR, CAR and lagged models);
- The distinction between a residual term and an error term, so that output for diagnosis of spatial autocorrelation and patterns in space contains four correlograms and maps (the

observed Y , the predicted Y , the residual and the error). If you use GLS, you will see that a strong pattern appears in the residuals, since you add this spatial element there! The error, on the other hand, must be spatially independent;

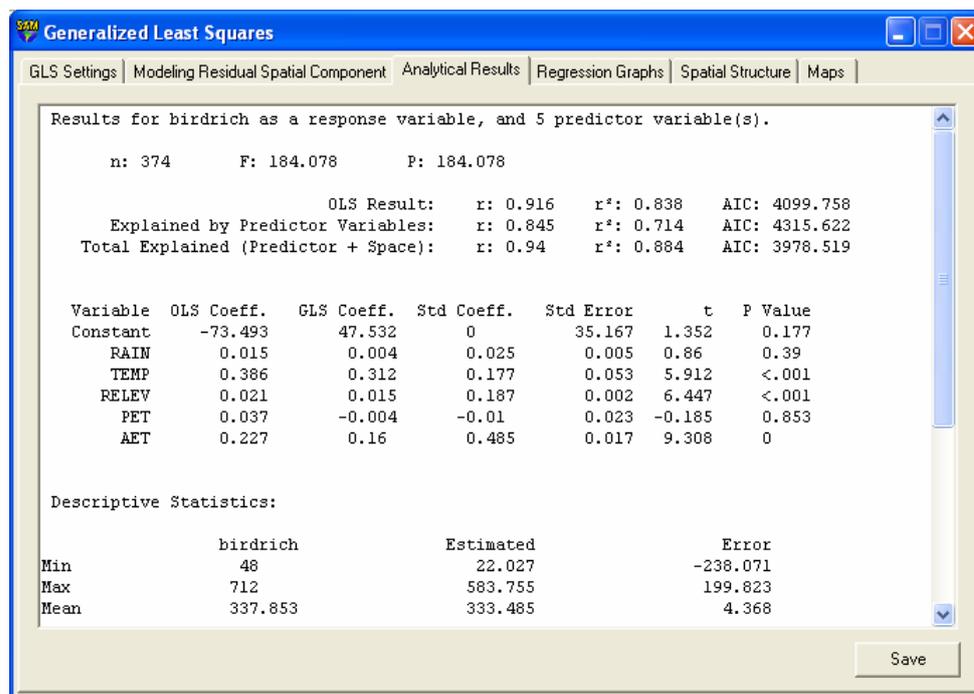
- For SAR and CAR, different R^2 values are given (understand their difference, and that you must be careful when interpreting them).

GLS or 'kriging' regression is very useful for learning about spatial modelling. The results below are from using all five environmental variables to model bird species richness. After clicking 'Next', you should see another screen, with a semi-variogram of an OLS model, and several modelling and diagnosis windows. These will help you to use different models (spherical, exponential, linear, and so on) with different parameters to describe the spatial structure in the OLS residuals. This is the same module as the one under Structure—Spatial autocorrelation.

We can try a spherical model, with C_0 equal to 1500, C_1 equal to 2650 and $a = 21$. A blue line describing this model appears. C_0 is the 'nugget' effect and describes the intercept of the line, whereas C_1 is the difference between the maximum expected semi-variance (the 'sill') and the intercept. The value ' a ' is the 'range' of the semi-variogram: the distance in which the sill is achieved. You can try different combinations of these three values (for other models as well), and check the diagnosis windows, in the lower right corner. In this window you have an OLS fit between the expected semi-variances (defined by your modelling process) and the observed ones, with the r^2 of the model. Using these values for nugget, sill-nugget and range, the spherical model has an r^2 of 0.829. The P-values associated with the t-tests are 0.027 (for the intercept) and 0.020 (for the slope). Thus, the results suggest that the chosen model is adequate for modelling the spatial structure in the residuals.



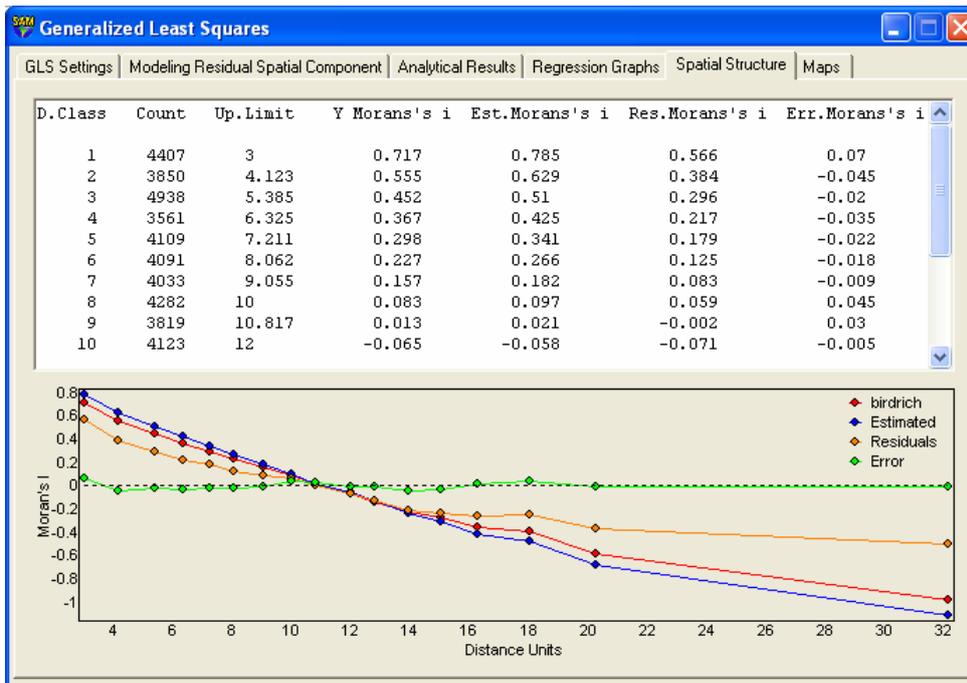
This is probably not the best possible model (you could check this later), but it seems to be good enough, so we can go on. Click 'Compute GLS' and you will have standard regression output, as before.



The main point now is that you have results from a standard OLS, for comparative purposes, and this includes the R² and AIC, in the first three main lines of the output.

The most awkward thing here is that you have two different R²s for GLS, and this requires some clarification. It is still debated whether GLS (as well as other spatial models) provides an R², in the sense of best fit, and what it really means. SAM uses the following strategy. The first R², defined as “Explained by predictors” and equal to 0.714 in this example, is the pseudo-R². This is estimated by correlating the observed values with the estimated values, which in turn were calculated using the partial regression slopes (shown below). This is smaller than the one from OLS because part of the spatial variation in richness that was ‘explained’ by the predictors in OLS was ‘transferred’ to the residuals in GLS, according to the semi-variance model you defined. However, this is not the ‘overall’ explanatory power of the model, because you added more information (i.e., the spatial structure). So another R² is calculated, using the ‘ratio’ of residual variance after fitting the full model and performing a Cholesky decomposition of spatially modelled residuals (you can check the HELP of SAM to get the mathematical detail). The important point is that this second R² is the overall explanation of the model, which includes both ‘predictors’ and space (in the residual spatial structure).

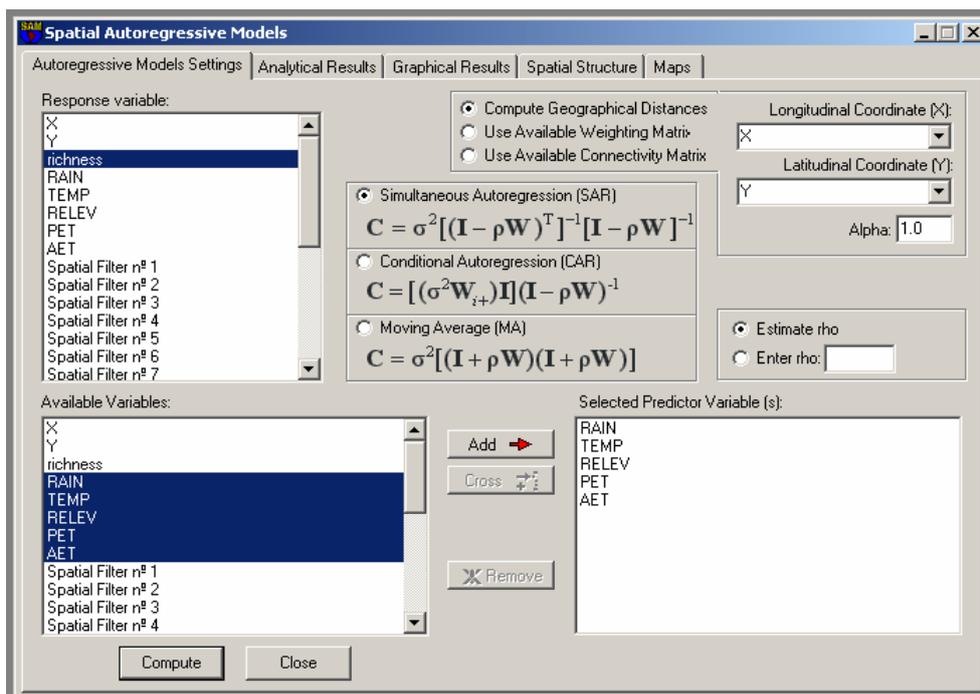
You can see the same thing in the “Spatial Structure” screen with the correlograms of the residuals. Here ‘residuals’ are defined as meaning the spatially modelled component of variation not accounted for by the environmental variables, and that’s why it has a very strong spatial structure (because you added this when modelling the semi-variogram!). By doing this you ‘released’, at the same time, the effect of the spatial structure of the environmental variables on richness. The ‘error’, on the other hand, is the non-spatially structured part of the error, and this should not have any spatial pattern. Indeed, this is what happens in the correlograms. However, the maps indicate that there is something about the tropical Andes that is still not being accounted for (notice that you now have four maps, instead of three).



Going back to the analytical results, examine the modelled effect of each environmental variable, and compare with the OLS results. In both, AET is the most 'important' environmental variable, according to the standardized slopes. This comparison of spatial and non-spatial solutions suggests some model stability. As occurred when using spatial filters in the model, RAIN is not significant now, but in the GLS the effect of PET is not significant either.

Exercise Mo9: Spatial autoregression – SAR and CAR

Press Ctrl-L or click  to bring up the Spatial Autoregressive Models dialog. Do the same thing as you have just done, but using SAR and CAR.



SAR and CAR are actually based on the same GLS routine; the only difference is that you are not allowed to model the residual autocorrelation in SAR and CAR, which is instead defined according to the equations shown in the screen and based on the same autoregressive parameter you calculated in the beginning of this exercise. Notice that, in SAR, for example, the ‘importance’ of environmental variables is similar to that in GLS: both PET and RAIN are not ‘significant’, although the P-values are close to the usual critical value of 0.05. There is also a relatively high spatial autocorrelation (Moran’s I) in the first distance class of errors, equal to 0.182 in the first distance class! This indicates that a better model should be tried, and you can do this by trying different definitions of spatial error structure, in the next activity.

Choose one autoregressive model (e.g., ‘pure’ autoregressive, SAR) and explore how using different alpha values (ranging from 1.0 to 3.0) and different definitions of spatial distances and connectivity affect the model outputs. Pay attention to R^2 , AIC, regression coefficients and residual and error autocorrelation.

To explore this, let’s go back to the ‘Pure Autoregressive Model’ (in the lagged models routine), for simplicity. If you change the alpha to 2.0, the weighting structure will be defined as $1/D^2$, so closer localities will be given more weight in generating the model. Doing this, the ‘Spatial Autoregressive Coefficient’ increases from 0.885 (using alpha = 1) to 0.917, and the Moran’s I in the first distance class goes down to 0.11. The AIC may be helpful, too: it decreases from 3123.92 to 2979.85, indicating a better model fit (differences higher than 3 can suggest improvement in the model). Going to alpha = 3, the autoregressive coefficient goes up further to 0.926, and AIC decreases to 2862.88. It is clear that increasing alpha improves model fit, and this tendency will continue for a while. In practice, alphas of 2 or 3 are usually considered enough to give a good description of fine-scale spatial structure.

If you use the rook connectivities (distances < 1.2), the autoregressive coefficient increases to 0.919 ± 0.007 , and the Moran’s I in the residuals is -0.023 (note that you do not have a correlogram for this, since you are using a connectivity matrix directly – but you can do one if you want, just save the residuals and go back to the ‘Spatial autocorrelation’ submodule).

Now go back to SAR and do the same thing, so you can check how changing spatial structure of errors affects the slopes of the environmental variables. You will see that the results converge to those from the GLS above, since AET is still the most ‘important’ environmental variable (but not with such a high relative importance as before). Using an alpha of 1 leaves some spatial autocorrelation structure in the residuals (0.182). With an alpha of 3, the Moran’s I of the error, in the first distance class, drops from 0.182 to 0.052. When using rook connectivities ($D < 1.2$), only RAIN is not significant, but it should be noted that the relative importance of AET is much reduced.

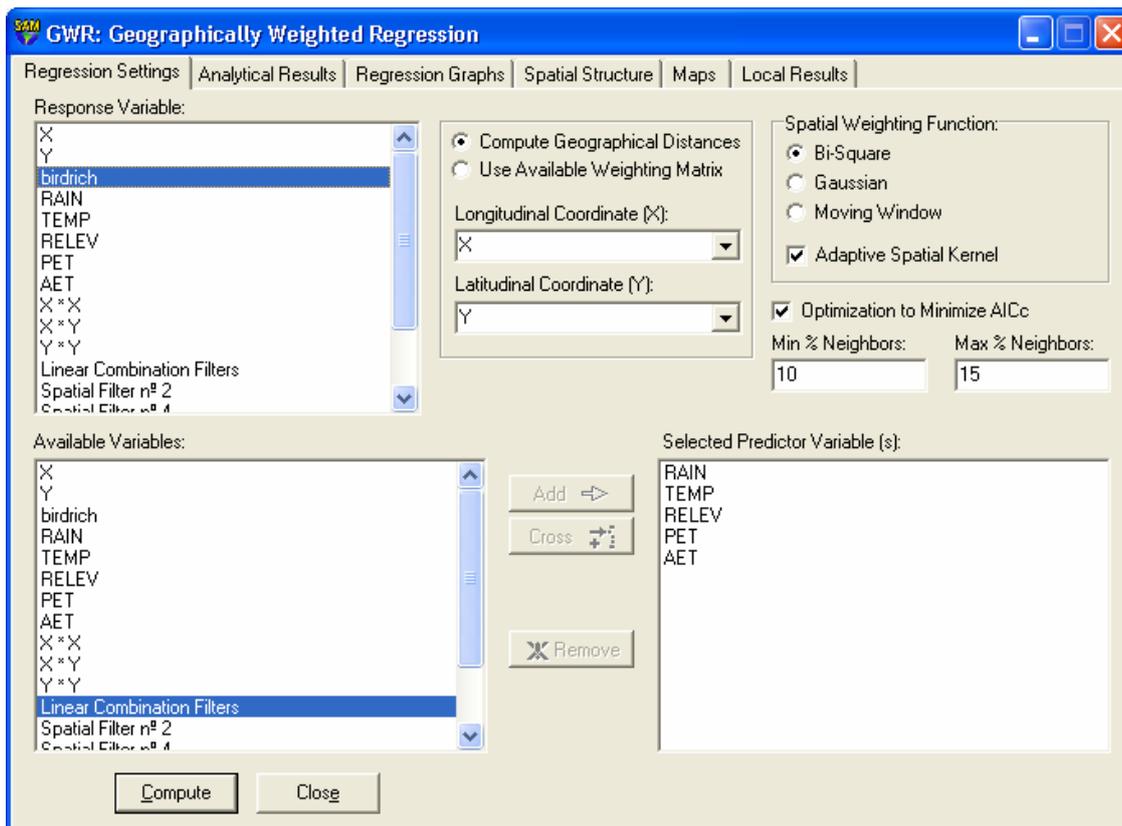
In this example of South American bird species richness, all spatial regression models show similar results, at least in the sense that AET is always the ‘best’ environmental ‘predictor’. The same applies for the OLS and filtering approaches. However, this will not always be the case, and each model should be carefully interpreted with respect to its assumptions, scale effects and robustness. To see a recent demonstration of exactly this, see the following paper, which a large group of us published recently:

Bini, L.M., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B. *et al.* (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. **Ecography** 32, 193–204.

Exercise Mo10: Geographically weighted regression

Geographically weighted regression (GWR) performs a series of local regressions, one for each cell in the grid. GWR is a relatively new method for local regression. The idea behind it is that there is a lot of information that could be extracted from the relationship between each cell and its neighbours. Thus, it is reasonable to perform one regression for each cell, taking into account the surrounding cells, but weighting them according to a distance function.

Press **Ctrl-W** or click  to bring up the **Geographically Weighted Regression dialog**. Specify the same regression model as before (in the example shown: the five environmental variables modelling bird species richness). In GWR you must, additionally, define a distance function (“Spatial Weighting Function” in the SAM dialog). Choose the options shown in the screen shot below and run the GWR.

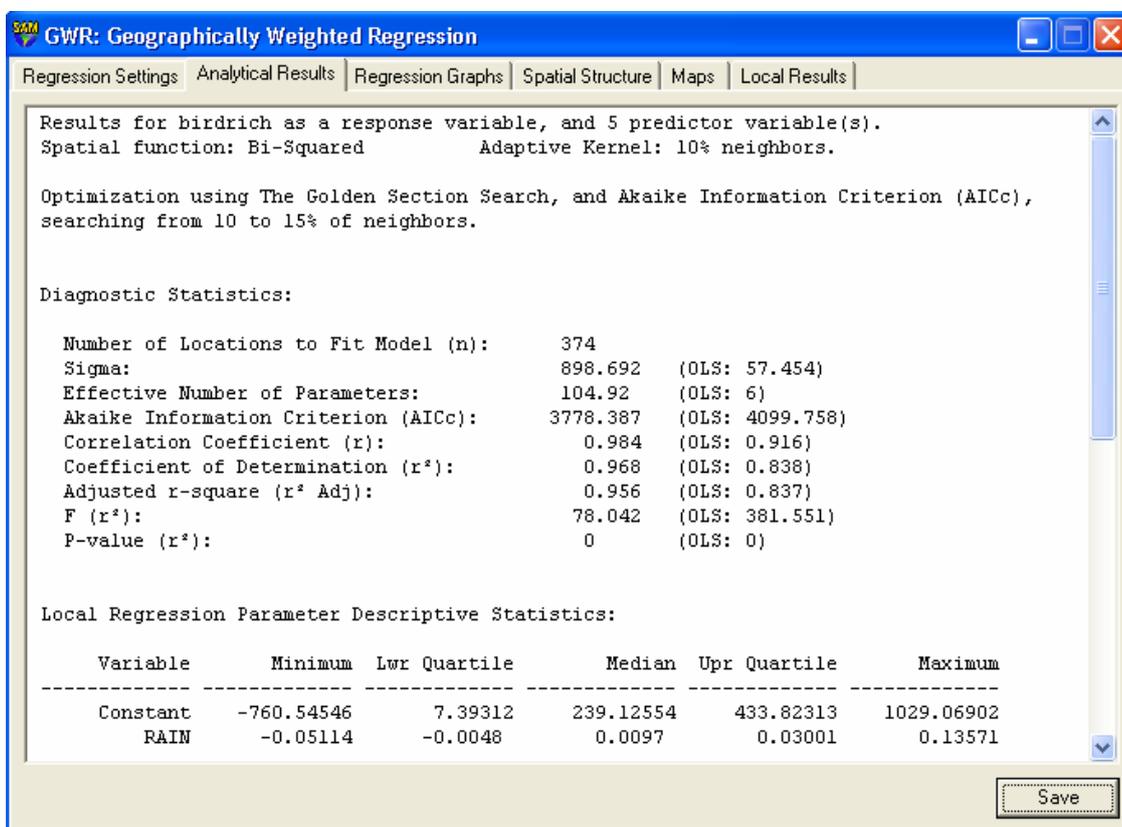


For “Spatial Weighting Function” in SAM, three options are available: “Bi-Square”, “Gaussian” and “Moving Window”. The first two are very similar, except that the “Bi-Square” truncates at a given distance (for this reason it is faster). Both draw a multi-dimensional normal distribution over the target cell, and weight the influence of the neighbours according to the normal distribution as a function of distance to the target cell. **This “normal distribution” has a parameter, the standard deviation, which we will call “bandwidth”**. Because it is a spatial

regression, the parameter “bandwidth” is measured in units of distance. This is a very important parameter in GWR, and defines what is “local” and “regional”.

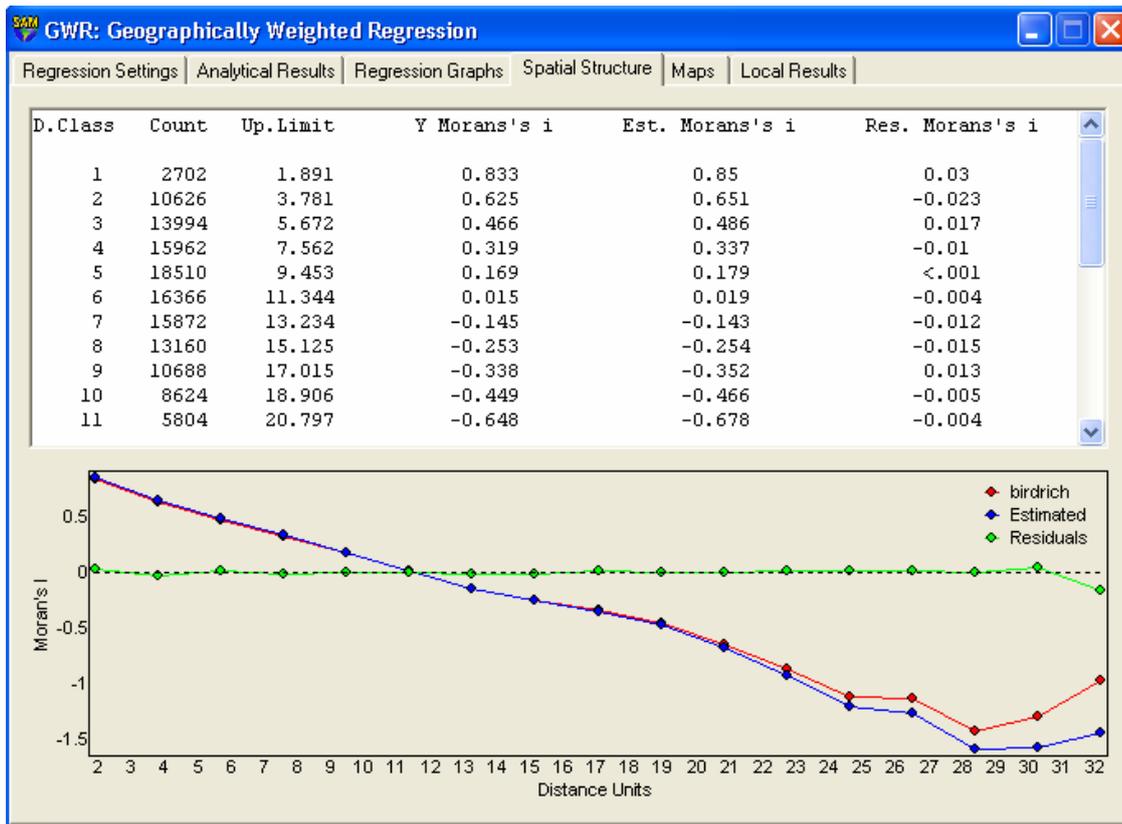
Another important option is the “Adaptive Spatial Kernel”. When this option is selected, instead of measuring the “bandwidth” in units of distance, you measure the “bandwidth” in units of neighbouring cells. This strategy helps avoiding border effects, and control for sampling effects among local regressions. Usually, this improves the fit of the model, and should be used if good fit is the aim.

Because the “bandwidth” is a parameter to be estimated (unless you have an specific measure of what is local and what is local), and it affects the goodness-of-fit of the model, it can be optimized. SAM searches for the bandwidth that minimizes the AICc if you select the option “Optimization to Minimize AICc”, and in this case you define the range of variation in the bandwidth to be searched.



In the Analytical Results tab you will find the GWR results and the OLS results (for comparison). The r^2 of the GWR is 0.968 (as opposed to 0.838 in OLS), and the AICc has decreased from 4100 in OLS to 3778 in GWR. However, to achieve such a close fit you lose generality, and now you have n linear equations to estimate the response variable (where n is the number of cases). The descriptive statistics of the parameters of those equations are shown in this same screen.

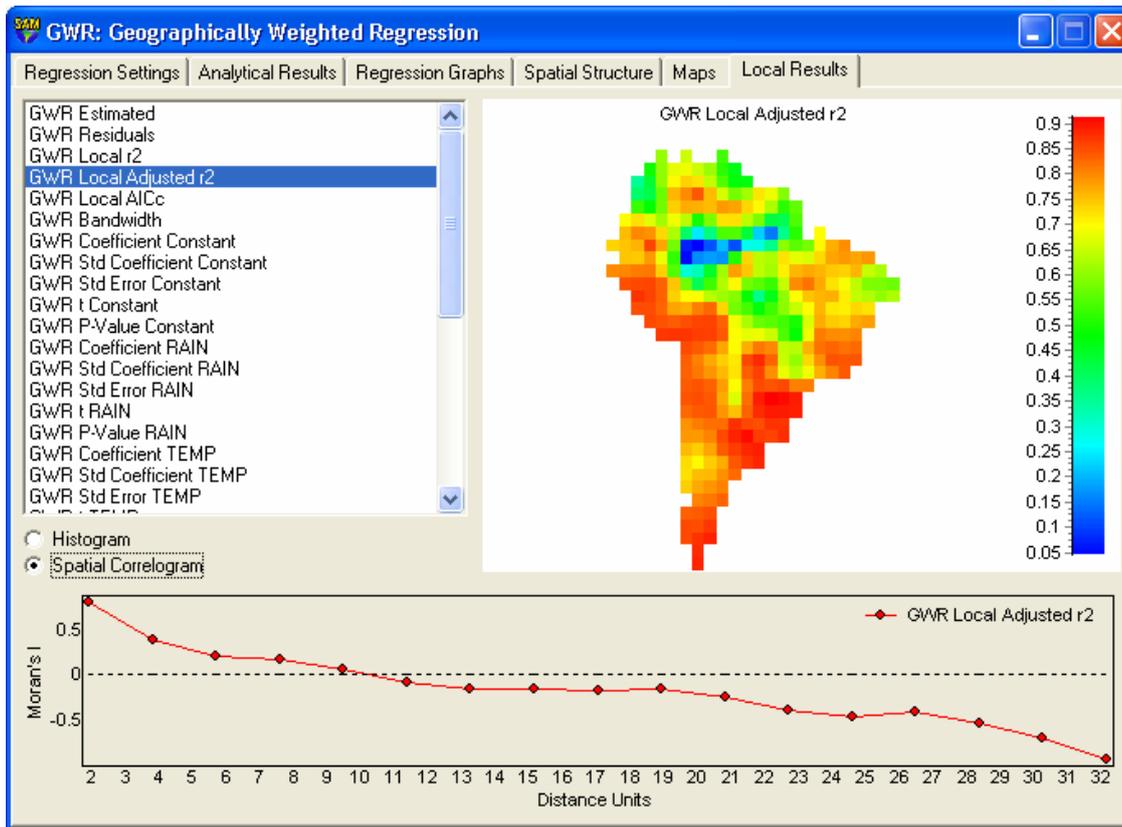
Also in the Analytical Results screen is an **ANOVA table to compare the explanatory power of the GWR with OLS**. It is very important to use this table, because GWR has 104.9 effective parameters (as opposed to only 6 in OLS), and thus it is expected provide more explanatory power compared with simpler models. This ANOVA table takes into account the over-parameterization of GWR, and is thus a fairer comparison with OLS.



As for the autocorrelation in the residuals, the correlogram shows that the model successfully accounted for all the contagious effect at short and long distances. You can try changing the bandwidth and see how it affects the correlogram of the residuals.

As expected, the map of the residuals is much more “random” than in the OLS, since the residuals are almost completely autocorrelation-free.

An additional result you have in GWR is the “Local Results”, in which you can examine the spatial variation in everything related to the regression! This is perhaps where the main interest lies, with GWR. Because GWR performs one regression in each cell, you also get all the standard OLS regression results per cell. In particular, you can map the r^2 or AICc (adjusted r^2 shown in the screen grab below) and check where the model could account for the local variation better and worse. Similarly, the maps of the regression coefficients, for each explanatory variable, can be useful for understanding the local fit of each variable throughout the map. All the variables available in this last tab can be saved for further analysis.



HELP

Finally, remember about the **HELP** file of the program, which can be double-clicked from Windows Explorer (in the Program Files) if it does not open from within SAM. Remember that SAM is being continuously updated and many new amendments are (always) under development. It is worth periodically checking the website (www.ecoevol.ufg.br/sam) to see/download these improvements. More importantly, please contact the SAM developers if you find bugs or problems, or have suggestions that can improve SAM.

We hope you have enjoyed this workshop, and that it has stimulated you. Perhaps many of your questions have been answered, but now replaced by much more complex ones! 😊

Best wishes!

Thiago (rangel.ufg@gmail.com), Richard and Alexandre.